

the loop of a conserved hairpin near the 3'-end of 16S rRNA. Inactivation of ksgA leads to resistance to the aminoglycoside antibiotic kasugamycin.

- Yeast 18S rRNA dimethylase (gene DIM1), which is functionally similar to ksgA and that dimethylates twin adenosines in the 3'-end of 18S rRNA.

- 5 - Bacterial 'erm' methylases. These enzymes confer resistance to macrolide-lincosamide-streptogramin B (MLS) antibiotics - such as erythromycin - by dimethylating the adenine residue at position 2058 of 23S rRNA thus resulting in a reduced affinity between ribosomes and the MLS antibiotics.
- *Caenorhabditis elegans* hypothetical protein EO2H1.1.

- 10 The best conserved regions in these enzymes is located in the N-terminal section and corresponds to a region that is probably involved in S-adenosyl methionine (SAM) binding.

- Consensus pattern: [LIVM]-[LIVMFY]-[DE]-x-G-[STAPV]-G-x-[GA]-x-[LIVMF]-[ST]-x(2)-[LIVM]-x(6)-[LIVMY]-x-[STAGV]-[LIVMFYIC]-E-x-D

- 15 [1] van Gemen B., van Knippenberg P.H.

(In) Nucleic acid methylation, Clawson G.A., Willis D.B., Weissbach A., Jones P.A., Eds., pp.19-36, Alan R. Liss Inc, New-York, (1990).

- [2] Lafontaine D., Delcour J., Glasser A.L., Desgres J., Vandenhoute J. J. Mol. Biol. 241:492-497(1994).

572. (RuBisCO small) Ribulose biphosphate carboxylase, small chain. 206 members

- 25 573. ATP/GTP-binding site motif A (P-loop) (ras)

From sequence comparisons and crystallographic data analysis it has been shown [1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP share a number of more or less conserved sequence motifs. The best conserved of these motifs is a glycine-rich region, which typically forms a flexible loop between a beta-strand and an alpha-helix.

- 30 This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is generally referred to as the 'A' consensus sequence [1] or the 'P-loop' [5]. There are numerous ATP- or GTP-binding proteins in which the P-loop is found. A number of protein families for which the relevance of the presence of such a motif has been noted are listed below: - ATP

synthase alpha and beta subunits. - Myosin heavy chains. - Kinesin heavy chains and kinesin-like proteins. - Dynamins and dynamin-like proteins - Guanylate kinase - Thymidine kinase (-Thymidylate kinase. - Shikimate kinase. - Nitrogenase iron protein family (nifH/irxC) - ATP-binding proteins involved in 'active transport' (ABC transporters) [7] - DNA and RNA
 5 helicases [8,9,10]. - GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc.). - Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.). - Nuclear protein ran. - ADP-ribosylation factors family - Bacterial dnaA protein - Bacterial recA protein - Bacterial recF protein - Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc.). - DNA mismatch repair proteins mutS family - Bacterial type II secretion system
 10 protein E. Not all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form: this is the case for tubulins or protein kinases. A special mention must
 15 be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.

Consensus pattern: [AG]-x(4)-G-K-[ST]

In addition to the proteins listed above, the 'A' motif is also found in a number of other proteins. Most of these proteins probably bind a nucleotide, but others are definitively not
 20 ATP- or GTP-binding (as for example chymotrypsin, or human ferritin light chain).

[1] Walker J.E., Saraste M., Runswick M.J., Gay N.J. EMBO J. 1:945-951(1982).[2] Moller W., Amons R. FEBS Lett. 186:1-7(1985).[3] Fry D.C., Kuby S.A., Mildvan A.S. Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).[4] Dever T.E., Glynias M.J., Merrick W.C. Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).[5] Saraste M., Sibbald P.R., Wittinghofer A. Trends
 25 Biochem. Sci. 15:430-434(1990).[6] Koonin E.V. J. Mol. Biol. 229:1165-1174(1993).[7] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J. Bioenerg. Biomembr. 22:571-592(1990).[8] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).[9] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P. Nature 337:121-122(1989).[10] Gorbalenya A.E.,
 30 Koonin E.V., Donchenko A.P., Blinov V.M. Nucleic Acids Res. 17:4713-4730(1989).

GTP-binding nuclear protein ran signature (ras)

Ran (or TC4) is a small abundant nuclear protein that binds and hydrolyzes GTP and which has been implicated in a large number of processes including nucleocytoplasmic transport, RNA synthesis, processing and export and cell cycle checkpoint control [1,2]. Ran is generally included in the RAS 'superfamily' of small GTP-binding proteins [3], but it is only slightly related to the other RAS proteins. It also differs from RAS proteins in that it lacks cysteine residues at its C- terminal and is therefore not subject to prenylation. Instead ran has an acidic C-terminus. It is, however similar to RAS family members in requiring a specific guanine nucleotide exchange factor (GEF) and a specific GTPase activating protein (GAP) as stimulators of overall GTPase activity. The region of the GTP-binding B motif which, in ran, is perfectly conserved has been selected as a signature pattern.

Consensus pattern: D-T-A-G-Q-E-K-[LF]-G-G-L-R-[DE]-G-Y-Y- Proteins belonging to this family also contain a copy of the ATP/GTP- binding motif 'A' (P-loop).

[1] Scheffzek K., Klebe C., Fritz-Wolf K., Kabseh W., Wittinghofer A. Nature 374:378-381(1995).[2] Rush M.G., Drivas G., d'Eustachio P. BioEssays 18:103-112(1996).[3] Valencia A., Chardin P., Wittinghofer A., Sander C. Biochemistry 30:4637-4648(1991).

574. recA signature

The bacterial recA protein [1,2,3,E1] is essential for homologous recombination and recombinational repair of DNA damage. RecA has many activities: it filaments, it binds to single- and double-stranded DNA, it binds and hydrolyzes ATP, it is also a recombinase and, finally, it interacts with lexA causing its activation and leading to its autocatalytic cleavage. RecA is a protein of about 350 amino-acid residues. Its sequence is very well conserved [3,4,5,E1] among eubacterial species. It is also found in the chloroplast of plants [6]. The best conserved region, a nonapeptide located in the middle of the sequence which is part of the monomer-monomer interface in a recA filament has been selected as a signature pattern.

Consensus pattern: A-L-[KR]-[IF]-[FY]-[STA]-[STAD]-[LIVMQ]-R-

[1] Smith K.C., Wang T.-C. V. BioEssays 10:12-16(1989).[2] Lloyd A.T., Sharp P.M. J. Mol. Evol. 37:399-407(1993).[3] Roca A.I., Cox M.M. Prog. Nucleic Acids Res. Mol. Biol. 56:129-223(1997).[4] Karlin S., Weinstock G.M., Brendel V. J. Bacteriol. 177:6881-6893(1995).[5] Eisen J.A. J. Mol. Evol. 41:1105-1123(1995).[6] Cerutti H.D., Osman M., Grandoni P., Jagendorf A.T. Proc. Natl. Acad. Sci. U.S.A. 89:8068-8072(1992).[E1]

<http://www.tigr.org/~jeisen/RecA/RecA.html>

575. Response regulator receiver domain

This domain receives the signal from the sensor partner inComment: bacterial two-component systems. It is usually found N-terminalComment: to a DNA binding effector domain.

[1] Pao GM, Saier MH; J Mol Evol 1995;40:136-154.

576. Ribonucleotide reductase large subunit signature

*Ribonucleotide reductase (EC 1.17.4.1) [1,2] catalyzes the reductive synthesis of deoxyribonucleotides from their corresponding ribonucleotides. It provides the precursors necessary for DNA synthesis. Ribonucleotide reductase is an oligomeric enzyme composed of a large subunit (700 to 1000 residues) and a small subunit (300 to 400 residues). There are regions of similarities in the sequence of the large chain from prokaryotes, eukaryotes and viruses. One of these regions has been selected as a signature pattern.

Consensus pattern: W-x(2)-[LF]-x(6,7)-G-[LIVM]-[FYRA]-[NH]-x(3)-[STAQLIVM]-[ASC]-x(2)-[PA]-

[1] Nillson O., Lundqvist T., Hahne S., Sjoberg B.-M. Biochem. Soc. Trans. 16:91-94(1988).[2] Reichard P. Science 260:1773-1777(1993).

577. Ribonuclease T2 family histidine active sites

The fungal ribonucleases T2 from *Aspergillus oryzae*, M from *Aspergillus saitoi* and Rh from *Rhizopus niveus* are structurally and functionally related 30 Kd glycoproteins [1] that cleave the 3'-5' internucleotide linkage of RNA via a nucleotide 2',3'-cyclic phosphate intermediates (EC 3.1.27.1). A number of other RNases have been found to be evolutionary related to these fungal enzymes: - Self-incompatibility [2] in flowering plants is often controlled by a single gene (S-gene) that has several alleles. This gene prevents fertilization by self-pollen or by pollen bearing either of the two S- alleles expressed in the style. The self-incompatibility glycoprotein from several higher plants of the solanaceae family has been shown [2,3] to be a ribonuclease. - Phosphate-starvation induced RNases LE and LX from tomato [4]. These two enzymes are probably involved in a phosphate-starvation rescue system. - *Escherichia coli*

periplasmic RNase I (EC 3.1.27.6) (gene rna) [5]. - *Aeromonas hydrophila* periplasmic RNase. - *Haemophilus influenzae* hypothetical protein HI0526. Two histidines residues have been shown [6,7] to be involved in the catalytic mechanism of RNase T2 and Rh. These residues and the region around them are highly conserved in all the sequence described above.

- 5 Two signature patterns have been developed, one for each of the two active-site histidines. The second pattern also contains a cysteine which is known to be involved in a disulfide bond.

Consensus pattern: [FYWL]-x-[LIVM]-H-G-L-W-P [H is an active site residue]

- 10 Consensus pattern: [LIVMF]-x(2)-[HDGTY]-[EQ]-[FYW]-x-[KR]-H-G-x-C [H is an active site residue] [C is involved in a disulfide bond]

[1] Watanabe H., Naitoh A., Suyama Y., Inokuchi N., Shimada H., Koyama T., Ohgi K., Irie M. J. Biochem. 108:303-310(1990).[2] Haring V., Gray J.E., McClure B.A., Anderson M.A., Clarke A.E. Science 250:937-941(1990).[3] McClure B.A., Haring V., Ebert P.R., Anderson M.A., Simpson R.J., Sakiyama F., Clarke A.E. Nature 342:95957(1989).[4] Loeffler A., Glund K., Irie M. Eur. J. Biochem. 214:627-633(1993).[5] Meador J. III, Kennell D. Gene 95:1-7(1990).[6] Kawata Y., Sakiyama F., Hayashi F., Kyogoku Y. Eur. J. Biochem. 187:255-262(1990).[7] Kurihara H., Mitsui Y., Ohgi K., Irie M., Mizuno H., Nakamura K.T. FEBS Lett. 306:189-192(1992).

578. Ribonucleotide reductase large subunit signature. Ribonucleotide reductase (EC 1.17.4.1) [1,2] catalyzes the reductive synthesis of deoxyribonucleotides from their corresponding ribonucleotides. It provides the precursors necessary for DNA synthesis. Ribonucleotide reductase is an oligomeric enzyme composed of a large subunit (700 to 1000 residues) and a small subunit (300 to 400 residues). There are regions of similarities in the sequence of the large chain from prokaryotes, eukaryotes and viruses. One of these regions has been developed as a signature pattern.

Consensus pattern: W-x(2)-[LF]-x(6,7)-G-[LIVM]-[FYRA]-[NH]-x(3)-[STAQLIVM]-[ASC]-x(2)-[PA]-

[1] Nillson O., Lundqvist T., Hahne S., Sjöberg B.-M. Biochem. Soc. Trans. 16:91-94(1988).[2] Reichard P. Science 260:1773-1777(1993).

579. RNase H

RNase H digests the RNA strand of an RNA/DNA hybrid. Important enzyme in retroviral replication cycle, and often found as a domain associated with reverse transcriptases. Structure is a mixed alpha+beta fold with three a/b/a layers.

580. Eukaryotic putative RNA-binding region RNP-1 signature (rrm)

Many eukaryotic proteins that are known or supposed to bind single-stranded RNA contain one or more copies of a putative RNA-binding domain of about 90 amino acids [1,2]. This region has been found in the following proteins: ** Heterogeneous nuclear ribonucleoproteins ** - hnRNP A1 (helix destabilizing protein) (twice). - hnRNP A2/B1 (twice). - hnRNP C (C1/C2) (once). - hnRNP E (UP2) (at least once). - hnRNP G (once). ** Small nuclear ribonucleoproteins ** - U1 snRNP 70 Kd (once). - U1 snRNP A (once). - U2 snRNP B" (once). ** Pre-RNA and mRNA associated proteins ** - Protein synthesis initiation factor 4B (eIF-4B) [3], a protein essential for the binding of mRNA to ribosomes (once). - Nucleolin (4 times). - Yeast single-stranded nucleic acid-binding protein (gene SSB1) (once). - Yeast protein NSR1 (twice). NSR1 is involved in pre-rRNA processing; it specifically binds nuclear localization sequences. - Poly(A) binding protein (PABP) (4 times). ** Others ** - Drosophila sex determination protein Sex-lethal (Sxl) (twice). - Drosophila sex determination protein Transformer-2 (Tra-2) (once). - Drosophila 'elav' protein (3 times), which is probably involved in the RNA metabolism of neurons. - Human paraneoplastic encephalomyelitis antigen HuD (3 times) [4], which is highly similar to elav and which may play a role in neuron-specific RNA processing. - Drosophila 'bicoid' protein (once) [5], a segment-polarity homeobox protein that may also bind to specific mRNAs. - La antigen (once), a protein which may play a role in the transcription of RNA polymerase III. - The 60 Kd Ro protein (once), a putative RNP complex protein. - A maize protein induced by abscisic acid in response to water stress, which seems to be a RNA-binding protein. - Three tobacco proteins, located in the chloroplast [6], which may be involved in splicing and/or processing of chloroplast RNAs (twice). - X16 [7], a mammalian protein which may be involved in RNA processing in relation with cellular proliferation and/or maturation. - Insulin-induced growth response protein Cl-4 from rat (twice). - Nucleolins TIA-1 and

TIAR (3 times) [8] which possesses nucleolytic activity against cytotoxic lymphocyte target cells. may be involved in apoptosis. - Yeast RNA15 protein, which plays a role in mRNA stability and/or poly-(A) tail length [9]. Inside the putative RNA-binding domain there are two regions which are highly conserved. The first one is a hydrophobic segment of six residues (which is called the RNP-2 motif), the second one is an octapeptide motif (which is called RNP-1 or RNP-CS). The position of both motifs in the domain is shown in the following schematic representation:

xxxxxxxx#####xx#####xx

RNP-2 RNP-1

The RNP-1 motif has been used as a signature pattern for this type of domain.

Consensus pattern: [RK]-G-{EDRKHPG}-[AGSCI]-[FY]-[LIVA]-x-[FYLM] In most cases the residue in position 3 of the pattern is either Tyr or Phe.

[1] Bandziulis R.J., Swanson M.S., Dreyfuss G. *Genes Dev.* 3:431-437(1989).[2] Dreyfuss G., Swanson M.S., Pinol-Roma S. *Trends Biochem. Sci.* 13:86-91(1988).[3] Milburn S.C., Hershey J.W.B., Davies M.V., Kelleher K., Kaufman R.J. *EMBO J.* 9:2783-2790(1990).[4] Szabo A., Dalmau J., Manley G., Rosenfeld M., Wong E., Henson J., Posner J.B., Furneaux H.M. *Cell* 67:325-333(1991).[5] Rebagliati M. *Cell* 58:231-232(1989).[6] Li Y., Sugiura M. *EMBO J.* 9:3059-3066(1990).[7] Ayane M., Preuss U., Koehler G., Nielsen P.J. *Nucleic Acids Res.* 19:1273-1278(1991).[8] Kawakami A., Tian Q., Duan X., Streuli M., Schlossman S.F., Anderson P. *Proc. Natl. Acad. Sci. U.S.A.* 89:8681-8685(1992).[9] Minvielle-Sebastia L., Winsor B., Bonneaud N., Lacroute F. *Mol. Cell. Biol.* 11:3075-3087(1991).

581. Rubredoxin signature

Rubredoxins [1] are small electron-transfer prokaryotic proteins. They contain an iron atom which is ligated by four cysteine residues. Rubredoxins are, in some cases, functionally interchangeable with ferredoxins.

A conserved region that includes two of the cysteine residues that bind the iron atom has been selected as a pattern for these proteins.

Consensus pattern: [LIVM]-x(3)-W-x-C-P-x-C-[AGD] [The two C's bind the iron atom]

In *Pseudomonas oleovorans* rubredoxin 2 (gene alkG) [2], this pattern is found twice because alkG has two rubredoxin domains.

Rubrerhythrin [3], a protein with inorganic pyrophosphatase activity from *Desulfovibrio vulgaris* possesses a C-terminal rubredoxin-like domain, but this domain is too divergent to be detected by the above pattern.

[1] Berg J.M., Holm R.H.(In) Iron-sulfur proteins, Spiro T.G., Ed., pp1-66, Wiley, New-York, (1982). [2] Kok M., Oldenhuis R., der Linden M.P.G., Meulenberg C.H.C., Kingma J., Witholt B., J. Biol. Chem. 264:5442-5451(1989). [3] van Beeumen J.J., van Driessche G., Liu M.-Y., Le Gall J., J. Biol. Chem. 266:20645-20653(1991).

582. (rvp) Eukaryotic and viral aspartyl proteases active site

Aspartyl proteases, also known as acid proteases, (EC 3.4.23.-) are a widely distributed family of proteolytic enzymes [1,2,3] known to exist invertebrates, fungi, plants, retroviruses and some plant viruses. Aspartate proteases of eukaryotes are monomeric enzymes which consist of two domains. Each domain contains an active site centered on a catalytic aspartyl residue. The two domains most probably evolved from the duplication of an ancestral gene encoding a primordial domain. Currently known eukaryotic aspartyl proteases are: - Vertebrate gastric pepsins A and C (also known as gastricsin). - Vertebrate chymosin (rennin), involved in digestion and used for making cheese. - Vertebrate lysosomal cathepsins D (EC 3.4.23.5) and E (EC 3.4.23.34). - Mammalian renin (EC 3.4.23.15) whose function is to generate angiotensin I from angiotensinogen in the plasma. - Fungal proteases such as aspergillopepsin A (EC 3.4.23.18), candidapepsin (EC 3.4.23.24), mucoropepsin (EC 3.4.23.23) (mucor rennin), endothiapepsin (EC 3.4.23.22), polyporopepsin (EC 3.4.23.29), and rhizopuspepsin (EC 3.4.23.21). - Yeast saccharopepsin (EC 3.4.23.25) (proteinase A) (gene PEP4). PEP4 is implicated in posttranslational regulation of vacuolar hydrolases. - Yeast barrier pepsin (EC 3.4.23.35) (gene BAR1); a protease that cleaves alpha-factor and thus acts as an antagonist of the mating pheromone. - Fission yeast sxa1 which is involved in degrading or processing the mating pheromones. Most retroviruses and some plant viruses, such as badnaviruses, encode for anaspartyl protease which is an homodimer of a chain of about 95 to 125 amino acids. In most retroviruses, the protease is encoded as a segment of a polyprotein which is cleaved during the maturation process of the virus. It is generally part of the pol polyprotein and, more rarely, of the gagpolyprotein. Conservation of the sequence

around the two aspartates of eukaryotic aspartyl proteases and around the single active site of the viral proteases allows us to develop a single signature pattern for both groups of protease.

Consensus pattern: [LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-
[STAPDENQ]-x-[LIVMFSTNC]-x-[LIVMFGTA] [D is the active site residue] –

- 5 [1] Foltmann B. Essays Biochem. 17:52-84(1981).[2] Davies D.R. Annu. Rev. Biophys. Chem. 19:189-215(1990).[3] Rao J.K.M., Erickson J.W., Wlodawer A. Biochemistry 30:4663-4671(1991).[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:105-120(1995).

10 583. (rvt) Reverse transcriptase (RNA-dependent DNA polymerase)

A reverse transcriptase gene is usually indicative of a mobile element such as a retrotransposon or retrovirus. Reverse transcriptases occur in a variety of mobile elements, including retrotransposons, retroviruses, group II introns, bacterial msDNAs, hepadnaviruses, and caulimoviruses. Number of members: 1233

[1] Medline: 91006031. Origin and evolution of retroelements based upon their reverse transcriptase sequences. Xiong Y, Eickbush TH; EMBO J 1990;9:3353-3362.

20 584. (S-AdoMet synt) S-adenosylmethionine synthetase signatures

S-adenosylmethionine synthetase (EC 2.5.1.6) is the enzyme that catalyzes the formation of S-adenosylmethionine (AdoMet) from methionine and ATP [1]. AdoMet is an important methyl donor for transmethylation and is also the propylamino donor in polyamine biosynthesis. In bacteria there is a single isoform of AdoMet synthetase (gene metK), there are two in budding yeast (genes SAM1 and SAM2) and in mammals while in plants there is generally a multigene family. The sequence of AdoMet synthetase is highly conserved throughout isozymes and species. Two signature patterns have been selected for this type of enzyme; the first is a hexapeptide which seems to be involved in ATP-binding; the second is an almost perfectly conserved glycine-rich nonapeptide.

25 30 Consensus pattern: G-A-G-D-Q-G-x(3)-G-[FYH]-Sequences known to belong to this class detected by the pattern:

Consensus pattern: G-[GA]-G-[ASC]-F-S-x-K-[DE]

[1] Horikawa S., Sasuga J., Shimizu K., Ozasa H., Tsukada K. J. Biol. Chem. 265:13683-13686(1990).

5 585. S1 RNA binding domain

The S1 domain occurs in a wide range of RNAComment: associated proteins. It is structurally similarComment: to cold shock protein which binds nucleic acids.Comment: The S1 domain has an OB-fold structure.

[1] Bycroft M, Hubbard TJ, Proctor M, Freund SM, Murzin AG; Cell 1997;88:235-242.

10

586. SAICAR synthetase signatures

Phosphoribosylaminoimidazole-succinocarboxamide synthase (EC 6.3.2.6)

(SAICARsynthetase) catalyzes the seventh step in the de novo purine biosynthetic pathway; the ATP-dependent conversion of 5'-phosphoribosyl-5-aminoimidazole-4-carboxylic acid and aspartic acid to SAICAR [1]. In bacteria (gene purC),fungi (gene ADE1) and plants, SAICAR synthetase is a monofunctional protein;in higher vertebrates it is the N-terminal domain of a bifunctional enzyme that also catalyze phosphoribosylaminoimidazole carboxylase (AIRC) activity. Two conserved regions in the central section of this enzyme have been selected as signature patterns for SAICAR synthetase.

Consensus pattern: [LIVMF](2)-P-[LIVM]-E-x-[LIVM]-[LIVMCA]-R-x(3)-[TA]-G-S-

Consensus pattern: [LIVM]-[LIVMA]-D-x-K-[LIVMFY]-E-F-G

[1] Zalkin H., Dixon J.E. Prog. Nucleic Acid Res. Mol. Biol. 42:259-287(1992).

25

587. (SCP) Extracellular proteins SCP/Tpx-1/Ag5/PR-1/Sc7 signatures

A variety of extracellular proteins from eukaryotes have been found to be evolutionary related: - Rodent sperm-coating glycoprotein (SCP), also known as acidic epididymal glycoprotein (AEG) . This protein is thought to be involved in sperm maturation [1]. It is a protein of about 220 residues and probably contains eight disulfide bonds. - Mammalian testis-specific protein Tpx-1 [2]. Tpx-1 is highly related to SCP's. - Mammalian glioma pathogenesis-related protein (GliPR). - Lizard helothermine, a toxin that blocks ryanodine receptors. - Venom allergen 5 (Ag5) from vespid wasps and venom allergen 3 (Ag3) from

30

fire ants. These proteins are potent allergens and are the main cause of allergic reactions to stings from insects of the hymenoptera family [3]. Ag5/3 are proteins of about 200 residues and contain four disulfide bonds. - Plant pathogenesis proteins of the PR-1 family [4]. These proteins are synthesized during pathogen infection or other stress-related responses. They are

5 proteins of about 130 to 140 residues and probably contain three disulfide bonds. - Proteins Sc7 and Sc14 from the basidiomycete fungus *Schizophyllum commune*. These extracellular proteins are loosely associated with fruit body hyphal walls [5]. Sc7/14 are proteins of about 180 residues and probably contain two disulfide bonds. - *Ancylostoma* secreted protein from dog hookworm. - Yeast hypothetical proteins YJL078c, YJL079c and YKR013w. The exact

10 function of these proteins is not yet known. Two conserved regions located in their C-terminal half have been selected as signature patterns. The second signature contains a cysteine which is known to be involved in a disulfide bond in Ag5.

Consensus pattern: [GDER]-H-[FYWH]-T-Q-[LIVM](2)-W-x(2)-[STN]

Consensus pattern: [LIVMFYH]-[LIVMFY]-x-C-[NQRHS]-Y-x-[PARH]-x-[GL]-N-[LIVMFYWDN] [C is involved in a disulfide bond]

[1] Mizuki N., Kasahara M. *Mol. Cell. Endocrinol.* 89:25-32(1992).[2] Kasahara M., Gutknecht J., Brew K., Spurr N., Goodfellow P.N. *Genomics* 5:527-534(1989).[3] Lu G., Villalba M., Coscia M.R., Hoffman D.R., King T.P. *J. Immunol.* 150:2823-2830(1993).[4] Dixon D.C., Cutt J.R., Klessig D.F. *EMBO J.* 10:1317-1324(1991).[5] Schuren F.H.J., Asgeirsdottir S.A., Kothe E.M., Scheer J.M.J., Wessels J.G.H. *J. Gen. Microbiol.* 139:2083-2090(1993).

588. SET domain

25 SET domains appear to be protein-protein interaction domains. It has been demonstrated that SET domains mediate interactions with a family of proteins that display similarity with dual-specificity phosphatases (dsPTPases) [2].

[1] Tripoulas N, LaJeunesse D, Gildea J, Shearn A; *Genetics* 1996;143:913-928. [2] Cui X, De Vivo I, Slany R, Miyamoto A, Firestein R, Cleary, ML; *Nat Genet* 1998;18:331-337.

589. Src homology 3 (SH3) domain profile

The Src homology 3 (SH3) domain is a small protein domain of about 60 amino-acid residues first identified as a conserved sequence in the non-catalytic part of several cytoplasmic protein tyrosine kinases (e.g. Src, Abl, Lck) [1]. Since then, it has been found in a great variety of other intracellular or membrane-associated proteins [2,3,4,5]. The SH3 domain has a characteristic fold which consists of five or six beta-strands arranged as two tightly packed anti-parallel beta sheets. The linker regions may contain short helices [6]. The function of the SH3 domain is not well understood. The current opinion is that they mediate assembly of specific protein complexes via binding to proline-rich peptides [7]. In general SH3 domains are found as single copies in a given protein, but there is a significant number of protein with two SH3 domains and a few with 3 or 4 copies. So far, SH3 domains have been identified in the following proteins: - Many vertebrate, invertebrate and retroviral cytoplasmic (non-receptor) protein tyrosine kinases. In particular in the Src, Abl, Bkt, Csk and ZAP70 families of kinases. - Mammalian phosphatidylinositol-specific phospholipase C-gamma-1 and -2. - Mammalian phosphatidyl inositol 3-kinase regulatory p85 subunit. - Mammalian Ras GTPase-activating protein (GAP). - Adaptor proteins mediating binding of guanine nucleotide exchange factors to growth factor receptors: vertebrate GRB2, *Caenorhabditis elegans* sem-5 and *Drosophila* DRK. All of which have two SH3 domains. - Mammalian Vav oncoprotein, a guanine nucleotide exchange factor of the CDC24 family. - Some guanine-nucleotide releasing factors of the CDC25 family: yeast CDC25, yeast SCD25, fission yeast ste6. - MAGUK proteins. These proteins consist of at least three types of domains: one or more copies of the DHR domain, a SH3 domain and a C-terminal guanylate kinase domain. Members of this family are: *Drosophila* lethal(1) discs large-1 tumor suppressor protein (gene Dlg1), mammalian tight junction protein ZO-1, vertebrate erythrocyte membrane protein p55, *Caenorhabditis elegans* protein lin-2, rat protein CASK and mammalian synaptic proteins SAP90/PSD-95, CHAPSYN-110/PSD-93, SAP97/DLG1 and SAP102. - Miscellaneous proteins interacting with vertebrate receptor protein tyrosine kinases: mammalian cytoplasmic protein Nck (3 copies), oncoprotein Crk (2 copies). - Chicken Src substrate p80/85 protein (cortactin) and the similar human hemopoietic lineage cell specific protein Hs1. - Mammalian dihydropyridine-sensitive L-type calcium channel beta (regulatory) subunit including the related human myasthenic syndrome antigen B (MSYB). - Mammalian neutrophil cytosolic activators of NADPH oxidase: p47 (NCF-1), p67 (NCF-2), and a potential homolog from *Caenorhabditis elegans* (B0303.7). NCF-1 and -2 have two copies of the SH3 domain, while B0303.7 has four. - Some myosin heavy chains from amoebae, slime

molds and yeast (gene MYO3). - Vertebrate and Drosophila spectrin and fodrin alpha-chain. - Human amphiphysin. - Yeast actin-binding protein ABP1. - Yeast actin-binding protein SLA1 (3 copies). - Yeast protein BEM1 and the fission yeast homolog scd2 (or ral3) (2 copies). - Yeast BEM1-binding proteins BOI2 (BEB1) and BOB1 (BOI1). - Yeast fusion protein FUS1. - Yeast protein RSV167. - Yeast protein SSU81. - Yeast hypothetical proteins YAR014c (1 copy), YFR024c (1 copy), YHL002w (1 copy), YHR016c (1 copy), YJL020C (1 copy), YHR114w (2 copies) and the fission yeast homolog SpAC12C2.05c. - *Caenorhabditis elegans* hypothetical proteins F42H10.3. The profile developed to detect SH3 domains is based on a structural alignment consisting of 5 gap-free blocks and 4 linker regions totaling 62 match positions.

[1] Mayer B.J., Hamaguchi M., Hanafusa H. *Nature* 332:272-275(1988).[2] Musacchio A., Gibson T., Lehto V.P., Saraste M. *FEBS Lett.* 307:55-61(1992).[3] Pawson T., Schlessinger J. *Curr. Biol.* 3:434-442(1993).[4] Mayer B.J., Baltimore D. *Trends Cell Biol.* 3:8-13(1993).[5] Pawson T. *Nature* 373:573-580(1995).[6] Kuriyan J., Cowburn D. *Curr. Opin. Struct. Biol.* 3:828-837(1993).[7] Morton C.J., Campbell I.D. *Curr. Biol.* 4:615-617(1994).

590. Serine hydroxymethyltransferase pyridoxal-phosphate attachment site (SHMT)

Serine hydroxymethyltransferase (EC 2.1.2.1) (SHMT) [1] catalyzes the transfer of the hydroxymethyl group of serine to tetrahydrofolate to form 5,10-methylenetetrahydrofolate and glycine. In vertebrates, it exists in acytoplasmic and a mitochondrial form whereas only one form is found in prokaryotes. Serine hydroxymethyltransferase is a pyridoxal-phosphate containing enzyme. The pyridoxal-P group is attached to a lysine residue around which the sequence is highly conserved in all forms of the enzyme.

Consensus pattern: [DEH]-[LIVMFY]-x-[STMV]-[GST]-[ST](2)-H-K-[ST]-[LF]-x-G-[PAC]-[RQ]-[GSA]-[GA] [K is the pyridoxal-P attachment site]

[1] Usha R., Savithri H.S., Rao N.A. *Biochim. Biophys. Acta* 1204:75-83(1994).

591. SIS domain

SIS (Sugar ISomerase) domains are found in many phosphosugar isomerases and phosphosugar binding proteins.

[1] Teplyakov A, Obmolova G, Badet-Denisot MA, Badet B, Polikarpov I; Structure 1998;6:1047-1055.

5 592. (SKI) Shikimate kinase signature

Shikimate kinase (EC 2.7.1.71) catalyzes the fifth step in the biosynthesis from chorismate of the aromatic amino acids (the shikimate pathway) in bacteria (gene *aroK* or *aroL*), plants and in fungi (where it is part of a multifunctional enzyme which catalyzes five consecutive steps in this pathway). Shikimate kinase is a small protein of about 200 residues. A conserved region that contains a run of three glycines has been selected as a signature pattern. Consensus pattern: [KR]-x(2)-E-x(3)-[LIVMF]-x(8,12)-[LIVMF](2)-[SA]-x-G(3)- x-[LIVMF]. Proteins belonging to this family also contain a copy of the ATP/GTP- binding motif 'A' (P-loop).

15 593. SNAP-25 family

SNAP-25 (synaptosome-associated protein 25 kDa) proteins are components of SNARE complexes. Members of this family contain a cluster of cysteine residues that can be palmitoylated for membrane attachment [2].

20 [1] Brennwald P, Kearns B, Champion K, Keranen S, Bankaitis V, Novick P; Cell 1994;79:245-258. [2] Risinger C, Blomqvist AG, Lundell I, Lambertsson A, Nassel D, Pieribone VA, Brodin L, Larhammar D; J Biol Chem 1993;268:24408-24414.

25 594. SNF2 and others N-terminal domain

This domain is found in proteins involved in a variety of processes including transcription regulation (e.g., SNF2, STH1, *brahma*, *MOT1*), DNA repair (e.g., ERCC6, RAD16, RAD5), DNA recombination (e.g., RAD54), and chromatin unwinding (e.g., ISWI) as well as a variety of other proteins with little functional information (e.g., *lodestar*, *ETL1*).

30

595. Staphylococcal nuclease homologues (Snase)

Present in all three domains of cellular life. Four copies in the transcriptional coactivator p100. These, however, appear to lack the active site residues of Staphylococcal nuclease.

Positions 14 (Asp-21), 34 (Arg-35), 39 (Asp-40), 42 (Glu-43) and Comment: 110 (Arg-87) [SNase numbering in parentheses] are thought to be involved in substrate-binding and catalysis.

[1] Ponting CP; Protein Sci 1997;6:459-463. [2] Callebaut I, Mornon JP; Biochem J 1997;321:125-132.

596. SPRY domainA

SPRY Domain is named from SPla and the RYanodine Receptor. Domain of unknown function. Distant homologues are domains in Comment: butyrophilin/marenostrin/pyrin homologues.

[1] Ponting C, Schultz J, Bork P; Trends Biochem Sci 1997;22:193-194.

597. (SQS PSY) Squalene and phytoene synthases signatures

Two different polyisoprene synthases have been shown [1,2,3] to share a number of regions of sequence similarities: - Squalene synthase (EC 2.5.1.21) (farnesyl-diphosphate farnesyltransferase) (SQS), which catalyzes the conversion of two molecules of farnesyl diphosphate (FPP) into squalene. It is the first committed step in the cholesterol biosynthetic pathway. The reaction carried out by SQS is catalyzed in two separate steps: the first is a head-to-head condensation of the two molecules of FPP to form presqualene diphosphate; this intermediate is then rearranged in a NADP-dependent reduction, to form squalene. SQS is found in eukaryotes. In yeast it is encoded by the ERG9 gene, in mammals by the FDFT1 gene. SQS seems to be membrane-bound. - Phytoene synthase (EC 2.5.1.-) (PSY), which catalyzes the conversion of two molecules of geranylgeranyl diphosphate (GGPP) into phytoene. It is the second step in the biosynthesis of carotenoids from isopentenyl diphosphate. The reaction carried out by PSY is catalyzed in two separate steps: the first is a head-to-head condensation of the two molecules of GGPP to form prephytoene diphosphate; this intermediate is then rearranged to form phytoene. PSY is found in all organisms that

synthesize carotenoids: plants and photosynthetic bacteria as well as some non-photosynthetic bacteria and fungi. In bacteria PSY is encoded by the gene crtB. In plants PSY is localized in the chloroplast. As it can be seen from the description above, both SQS and PSY share a number of functional similarities which are also reflected at the level of their primary structure. In particular three well conserved regions are shared by SQS and PSY; they could be involved in substrate binding and/or the catalytic mechanism. Signature patterns have been developed for the second and third conserved regions; they are localized in the central part of these enzymes.

Consensus pattern: Y-[CSAM]-x(2)-[VSG]-A-[GSA]-[LIVAT]-[IV]-G-x(2)-[LMSC]- x(2)-[LIV]

Consensus pattern: [LIVM]-G-x(3)-Q-x(2,3)-N-[IF]-x-R-D-[LIVMFY]-x(2)-[DE]- x(4,7)-R-x-[FY]-x-P-

[1] Summers C., Karst F., Charles A.D. Gene 136:185-192(1993).[2] Robinson G.W., Tsay Y.H., Kienzle B.K., Smith-Monroy C.A., Bishop R.W. Mol. Cell. Biol. 13:2706-2727(1993).[3] Roemer S., Huguency P., Bouvier F., Camara B., Kuntz M. Biochem. Biophys. Res. Commun. 196:1414-1421(1993).

598. SRP54-type proteins GTP-binding domain signature

The signal recognition particle (SRP) is an oligomeric complex that mediates targeting and insertion of the signal sequence of exported proteins into the membrane of the endoplasmic reticulum. SRP consists of a 7S RNA and six protein subunits. One of these subunits, the 54 Kd protein (SRP54), is a GTP-binding protein that interacts with the signal sequence when it emerges from the ribosome. The N-terminal 300 residues of SRP54 include the GTP-binding site (G-domain) and are evolutionary related to similar domains in other proteins which are listed below [1]. - Escherichia coli and Bacillus subtilis ffh protein (P48), a protein which seems to be the prokaryotic counterpart of SRP54. Ffh is associated with a 4.5S RNA in the prokaryotic SRP complex. - Signal recognition particle receptor alpha subunit (docking protein), an integral membrane GTP-binding protein which ensures, in conjunction with SRP, the correct targeting of nascent secretory proteins to the endoplasmic reticulum membrane. The G-domain is located at the C-terminal extremity of the protein. - Bacterial ftsY protein, a protein which is believed to play a similar role to that of the docking protein in eukaryotes. The G-domain is located at the C-terminal extremity of the protein. - The pilA protein from

Neisseria gonorrhoeae which seems to be the homolog of ftsY. - A protein from the archaeobacteria Sulfolobus solfataricus. This protein is also believed to be a docking protein. The G-domain is also at the C- terminus. - Bacterial flagellar biosynthesis protein flhF. The best conserved regions in those domains are the sequence motifs that are part of the GTP-binding site, but as those regions are not specific to these proteins, they were not used as a signature pattern. Instead, a conserved region located at the C-terminal end of the domain was selected.

Consensus pattern: P-[LIVM]-x-[FYI]-[LIVMAT]-[GS]-x-[GS]-[EQ]-x(4)-[LIVMF]
[1] Althoff S., Selinger D., Wise J.A. Nucleic Acids Res. 22:1933-1947(1994).

599. (STphosphatase) Serine/threonine specific protein phosphatases signature
Serine/threonine specific protein phosphatases (EC 3.1.3.16) (PP) [1,2,3] are enzymes that catalyze the removal of a phosphate group attached to a serine or evolutionary related. -
Protein phosphatase-1 (PP1) is an enzyme of broad specificity. It is inhibited by two thermostable proteins, inhibitor-1 and -2. In mammals, there are two closely related isoforms of PP-1: PP-1alpha and PP-1beta, produced by alternative splicing of the same gene. In Emericella nidulans, PP-1 (gene bimG) plays an important role in mitosis control by reversing the action of the nimA kinase. In yeast, PP-1 (gene SIT4) is involved in dephosphorylating the large subunit of RNA polymerase II. - Protein phosphatase-2A (PP2A) is also an enzyme of broad specificity. PP2A is a trimeric enzyme that consist of a core composed of a catalytic subunit associated with a 65 Kd regulatory subunit and a third variable subunit. In mammals, there are two closely related isoforms of the catalytic subunit of PP2A: PP2A-alpha and PP2A-beta, encoded by separate genes. - Protein phosphatase-2B (PP2B or calcineurin), a calcium-dependent enzyme whose activity is stimulated by calmodulin. It is composed of two subunits: the catalytic A-subunit and the calcium-binding B-subunit. The specificity of PP2B is restricted. In addition to the above-mentioned enzymes, some additional serine/threoninespecific protein phosphatases have been characterized and are listed below. - Mammalian phosphatase-X (PP-X), and Drosophila phosphatase-V (PP-V) which are closely related but yet distinct from PP2A. - Yeast phosphatase PPH3, which is similar to PP2A, but with different enzymatic properties. - Drosophila phosphatase-Y (PP-Y), and yeast phosphatases Z1 and Z2 (genes PPZ1 and PPZ2) which are closely related but yet distinct from PP1. - Drosophila retinal degeneration protein C (gene rdgC), a calcium-binding

phosphatase required to prevent light-induced retinal degeneration. - Phages Lambda and Phi-80 ORF-221 which have been shown to have phosphatase activity and are related to mammalian PP's. The best conserved regions in these proteins is a perfectly conserved pentapeptide that can be used as a signature pattern.

5 Consensus pattern: [LIVM]-R-G-N-H-E-

[1] Cohen P. Annu. Rev. Biochem. 58:453-508(1989).[2] Cohen P., Cohen P.T.W. J. Biol. Chem. 264:21435-21438(1989).[3] Cohen P.T.W., Brewis N.D., Hughes V., Mann D.J. FEBS Lett. 268:355-359(1990).

10

600. Translation initiation factor SUI1 signature

In budding yeast (*Saccharomyces cerevisiae*), SUI1 is a translation initiation factor that functions in concert with eIF-2 and the initiator tRNA-Met in directing the ribosome to the proper start site of translation [1]. SUI1 is a protein of 108 residues. Close homologs of SUI1 have been found [2] in mammals, insects and plants. SUI1 is also evolutionary related to hypothetical proteins from *Escherichia coli* (*yciH*), *Haemophilus influenzae* (HI1225) and *Methanococcus vannielii*. A conserved region in the C-terminal section has been selected as a signature pattern.

Consensus pattern: [LIVM]-[EQ]-[LIVM]-Q-G-[DEN]-[KHQ]-[KRV]

[1] Yoon H., Donahue T.F. Mol. Cell. Biol. 12:248-260(1992).[2] Fields C.A., Adams M.D. Biochem. Biophys. Res. Commun. 198:288-291(1994).

601. (S T dehydratase) Serine/threonine dehydratases pyridoxal-phosphate attachment site

25 Serine and threonine dehydratases [1,2] are functionally and structurally related pyridoxal-phosphate dependent enzymes: - L-serine dehydratase (EC 4.2.1.13) and D-serine dehydratase (EC 4.2.1.14) catalyze the dehydration of L-serine (respectively D-serine) into ammonia and pyruvate. - Threonine dehydratase (EC 4.2.1.16) (TDH) catalyzes the dehydration of threonine into alpha-ketobutarate and ammonia. In *Escherichia coli* and
30 other microorganisms, two classes of TDH are known to exist. One is involved in the biosynthesis of isoleucine, the other in hydroxamino acid catabolism. Threonine synthase (EC 4.2.99.2) is also a pyridoxal-phosphate enzyme, it catalyzes the transformation of homoserine-phosphate into threonine. It has been shown [3] that threonine synthase is

distantly related to the serine/threonine dehydratases. In all these enzymes, the pyridoxal-phosphate group is attached to a lysine residue. The sequence around this residue is sufficiently conserved to allow the derivation of a pattern specific to serine/threonine dehydratases and threonine synthases.

5 Consensus pattern: [DESH]-x(4,5)-[STVG]-x-[AS]-[FYI]-K-[DLIFSA]-[RVMF]-[GA]-[LIVMGA] [The K is the pyridoxal-P attachment site]

[1] Ogawa H., Gomi T., Konishi K., Date T., Naakashima H., Nose K., Matsuda Y., Peraino C., Pitot H.C., Fujioka M. J. Biol. Chem. 264:15818-15823(1989).[2] Datta P., Goss T.J., Omnaas J.R., Patil R.V. Proc. Natl. Acad. Sci. U.S.A. 84:393-397(1987).[3] Parsot C. EMBO J. 5:3013-3019(1986).[4] Grabowski R., Hofmeister A.E.M., Buckel W. Trends Biochem. Sci. 18:297-300(1993).

Cysteine synthase/cystathionine beta-synthase P-phosphate attachment site

Cysteine synthase (CSase) is the pyridoxal-phosphate dependent enzyme responsible [1] for the formation of cysteine from O-acetyl-serine and hydrogen sulfide with the concomitant release of acetic acid. In bacteria such as Escherichia coli, two forms of the enzyme are known (genes *cysK* and *cysM*). In plants there are also two forms, one located in the cytoplasm and the other in chloroplasts. Cystathionine beta-synthase [2] catalyzes the first irreversible step in homocysteine transsulfuration; the conjugation of homocysteine and serine forming cystathionine. Like CSase it is a pyridoxal-phosphate dependent enzyme. The two types of enzymes are evolutionary related. The pyridoxal-phosphate group of CSases has been shown to be attached to a lysine residue which is located in the N-terminal section of these enzymes; the sequence around this residue is highly conserved and can be used as a signature pattern to detect this class of enzymes.

25 Consensus pattern: K-x-E-x(3)-[PA]-[STAGC]-x-S-[IVAP]-K-x-R-x-[STAG]-x(2)- [LIVM] [The 2nd K is the pyridoxal-P attachment site]

[1] Saito K., Kurosawa M., Murakoshi I. FEBS Lett. 328:111-114(1993).[2] Swaroop M., Bradley K., Ohura T., Tahara T., Roper M.D., Rosenberg L.E., Kraus J.P. J. Biol. Chem. 267:11455-11461(1992).

30

602. S locus glycop

S-locus glycoprotein family. In Brassicaceae, self-incompatible plants have a self/non-self
 Comment: recognition system. This is sporophytically controlled by Comment: multiple
 alleles at a single locus (S). S-locus glycoproteins, Comment: as well as S-receptor kinases,
 are in linkage with the S-alleles [1]. Number of members: 128

[1] Evolutionary aspects of the S-related genes of the Brassica self-incompatibility system:
 synonymous and nonsynonymous base substitutions. Hinata K, Watanabe M, Yamakawa S,
 Satta Y, Isogai A; Genetics 1995;140:1099-1104. [2] Polymorphism of the S-locus
 glycoprotein gene (SLG) and the S-locus related gene (SLR1) in *Raphanus sativus* L. and
 self-incompatible ornamental plants in the Brassicaceae. Sakamoto K, Kusaba M, Nishio T;
 Mol Gen Genet 1998;258:397-403.

603. (sdh cyt) Succinate dehydrogenase cytochrome b subunit signatures

Succinate dehydrogenase (SDH) is a membrane-bound complex of two main components: a
 membrane-extrinsic component composed of an FAD-binding flavoprotein and an iron-sulfur
 protein, and a hydrophobic component composed of a cytochrome B and a membrane anchor
 protein. The cytochrome b component is a mono heme transmembrane protein [1,2,3]
 belonging to a family that groups: - Cytochrome b-556 from bacterial SDH (gene *sdhC*). -
 Cytochrome b560 from the mammalian mitochondrial SDH complex. - Cytochrome b560
 subunit encoded in the mitochondrial genome of some algae and in the plant *Marchantia*
polymorpha. - Cytochrome b from yeast mitochondrial SDH complex (gene *SDH3* or *CYB3*).
 - Protein cyt-1 from *Caenorhabditis*. These cytochromes are proteins of about 130 residues
 that comprise three transmembrane regions. There are two conserved histidines which may
 be involved in binding the heme group. Two signature patterns have been developed that
 include these histidine residues.

Consensus pattern: R-P-[LIVMT]-x(3)-[LIVM]-x(6)-[LIVMWPK]-x(4)-S-x(2)-H-R-x- [ST]
 [H could be a heme ligand]

Consensus pattern: H-x(3)-[GA]-[LIVMT]-R-[HF]-[LIVMF]-x-[FYWM]-D-x-[GVA] [H
 could be a heme ligand]

[1] Yu L., Wei Y.-Y., Usui S., Yu C.-A. J. Biol. Chem. 267:24508-24515(1992).[2]
 Abraham P.R., Mulder A., Van't Riet J., Raue H.A. Mol. Gen. Genet. 242:708-716(1994).[3]
 Leblanc C., Boyen C., Richard O., Bonnard G., Grienemberger J.M., Kloareg B. J. Mol. Biol.
 250:484-495(1995).

604. Sec1 family

[1] The Sec1 family: a novel family of proteins involved in synaptic transmission and general secretion. Halachmi N, Lev Z; J Neurochem 1996;66:889-897.

Number of members: 40

605. Protein secE/sec61-gamma signature

In bacteria, the secE protein plays a role in protein export; it is one of the components - with secY and secA - of the preprotein translocase. In eukaryotes, the evolutionary related protein sec61-gamma plays a role in protein translocation through the endoplasmic reticulum; it is part of a trimeric complex that also consists of sec61-alpha and beta [1]. Both secE and sec61-gamma are small proteins of about 60 to 90 amino acids that contain a single transmembrane region at their C-terminal extremity (Escherichia coli secE is an exception, in that it possesses an extra N-terminal segment of 60 residues that contains two additional transmembrane domains). The sequence of secE/sec61-gamma is not extremely well conserved, however it is possible to derive a signature pattern centered on a conserved proline located 10 residues before the beginning of the transmembrane domain.

Consensus pattern: [LIVMFY]-x(2)-[DENQGA]-x(4)-[LIVMFTA]-x-[KRV]-x(2)-[KW]-P-x(3)-[SEQ]-x(7)-[LIVT]-[LIVGA]-[LIVFGAST]

[1] Hartmann E., Sommer T., Prehn S., Goerlich D., Jentsch S., Rapoport T.A. Nature 367:654-657(1994).

606. 11-S plant seed storage proteins signature

Plant seed storage proteins, whose principal function appears to be the major nitrogen source for the developing plant, can be classified, on the basis of their structure, into different families. 11-S are non-glycosylated proteins which form hexameric structures [1,2]. Each of the subunits in the hexamer is itself composed of an acidic and a basic chain derived from a single precursor and linked by a disulfide bond. This structure is shown in the following representation. +-----+ ||

xxxxxxxxxxCxxxxxxxxxxxxxxxxxxxxxxxxxxNGxCxxxxxxxxxxxxxxxxxxxxxxxxxx ***** <--

----Acidic-subunit-----><----Basic-subunit-----> <-----About-480-to-500-residues----->'C': conserved cysteine involved in a disulfide bond.'*': position of the pattern. Proteins that belong to the 11-S family are: pea and broad bean legumins, rape cruciferin, rice glutelins, cotton beta-globulins, soybean glycinins, pumpkin 11-S globulin, oat globulin, sunflower helianthinin G3, etc. The region that includes the conserved cleavage site between the acidic and basic subunits (Asn-Gly) and a proximal cysteine residue which is involved in the interchain disulfide bond have been used as a signature pattern for this family of proteins.

Consensus pattern: N-G-x-[DE](2)-x-[LIVMF]-C-[ST]-x(11,12)-[PAG]-D [C is involved in a disulfide bond

[1] Hayashi M., Mori H., Nishimura M., Akazawa T., Hara-Nishimura I. Eur. J. Biochem. 172:627-632(1988).[2] Shotwell M.A., Afonso C., Davies E., Chesnut R.S., Larkins B.A. Plant Physiol. 87:698-704(1988).

607. 7S seed storage protein

7S globulin is one of the main storage proteins of most angiosperms and gymnosperms. The 7S storage proteins are homotrimers.

Number of members: 67

[1] The three-dimensional structure of canavalin from jack bean (*Canavalia ensiformis*). Ko TP, Ng JD, McPherson A; Plant Physiol 1993;101:729-744.

608. Aspartate-semialdehyde dehydrogenase signature

Aspartate-semialdehyde dehydrogenase (ASD) catalyzes the second step in the common biosynthetic pathway leading from Asp to diaminopimelate and Lys, to Met, and to Thr; the NADP-dependent reductive dephosphorylation of L-aspartyl phosphate to L-aspartate-semialdehyde. In bacteria and fungi, ASD is a protein of about 40 Kd (340 to 370 residues) whose sequence is not extremely well conserved [1]. A conserved cysteine residue has been implicated as important for the catalytic activity [2]. The region of conservation around the active site residue is too small to be used as signature pattern. Another more conserved region, located in the last third of the sequence, and which contains both a conserved cysteine as well as an histidine has been used instead.

Consensus pattern: [LIVM]-[SADN]-x(2)-C-x-R-[LIVM]-x(4)-[GSC]-H-[STA

[1] Baril C., Richaud C., Fourni E., Baranton G., Saint Girons I. J. Gen. Microbiol. 138:47-53(1992).[2] Karsten W.E., Viola R.E. Biochim. Biophys. Acta 1121:234-238(1992).

5 N-acetyl-gamma-glutamyl-phosphate reductase active site

N-acetyl-gamma-glutamyl-phosphate reductase (EC 1.2.1.38) (AGPR) [1,2] is the enzyme that catalyzes the third step in the biosynthesis of arginine from glutamate, the NADP-dependent reduction of N-acetyl-5-glutamyl phosphate into N-acetylglutamate 5-semialdehyde. In bacteria it is a monofunctional protein of 35 to 38 Kd (gene argC) while in
10 fungi it is part of a bifunctional mitochondrial enzyme (gene ARG5,6, arg11 or arg-6) which contains a N-terminal acetylglutamate kinase (EC 2.7.2.8) domain and a C-terminal AGPR domain. In the Escherichia coli enzyme, a cysteine has been shown to be implicated in the catalytic activity, the region around this residue is well conserved and can be used as a signature pattern.

15 Consensus pattern: [LIVM]-[GSA]-x-P-G-C-[FY]-[AVP]-T-[GA]-x(3)-[GTAC]-[LIVM]- x-P [C is the active site residue]

[1] Ludovice M., Martin J.F., Carrachas P., Liras P. J. Bacteriol. 174:4606-4613(1992).[2] Gessert S.F., Kim J.H., Nargang F.E., Weiss R.L. J. Biol. Chem. 269:8189-8203(1994).

20 609. Sialyltransferase family,

Number of members: 18

25 610. SpoU rRNA Methylase family

This family of proteins probably use S-AdoMet. Number of members: 58

[1] SpoU protein of Escherichia coli belongs to a new family of putative rRNA methylases.

Koonin EV, Rudd KE; Nucleic Acids Res 1993;21:5519-5519. [2] The spoU gene of escherichia coli , the fourth gene of the spoT operon, is essential for tRNA (Gm18) 2 '

30 methyltransferase activity. Persson BC, Jager G, Gustafsson C; Nucleic Acids Res 1997;25:4093-4097.

611. Stathmin family signatures

Stathmin [1] (from the Greek 'stathmos' which means relay), is an ubiquitous intracellular protein, present in a variety of phosphorylated forms and which serves as a relay for diverse second messenger pathways. Its expression and phosphorylation are regulated throughout development and in response to extracellular signals regulating cell proliferation, differentiation and function. Stathmin is a highly conserved protein of 149 amino acid residues. Structurally, it consists of an N-terminal domain of about 45 residues followed by a 78 residue alpha-helical domain consisting of a heptad repeat coiled coil structure and a C-terminal domain of 25 residues. Protein SCG10 is a neuron-specific, membrane-associated protein that accumulates in the growth cones of developing neurons. It is highly similar in its sequence to stathmin, but differs in that it contains an additional N-terminal hydrophobic segment of 32 residues which is probably responsible for its interaction with membranes. Xenopus protein XB3 is also evolutionary related to stathmin and also contains an additional N-terminal hydrophobic domain [2]. A conserved decapeptide which ends with the first three residues of the coiled coil domain and a second pattern that corresponds to part of the central region of the coiled coil have been selected as signatures for proteins of the stathmin family. Consensus pattern: P-[KRQ]-[KR](2)-[DE]-x-S-L-[EG]-E- Consensus pattern: A-E-K-R-E-H-E-[KR]-E- [1] Sobel A. Trends Biochem. Sci. 16:301-305(1991).[2] Maucuer A., Moreau J., Mechali M., Sobel A. J. Biol. Chem. 268:16420-16429(1993).

612. SUA5/ydcO/yrdC family signature. The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast protein SUA5. - Escherichia coli hypothetical protein ydcO and HI1198, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yrdC and HI0656, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein ywIC. - Mycobacterium leprae hypothetical protein in rfe-hemK intergenic region. - Methanococcus jannaschii hypothetical protein MJ0062. These are proteins of from 20 to 46 Kd which contain a number of conserved regions in their N-terminal section. They can be picked up in the database by the following pattern.

Consensus pattern: [LIVMTA](3)-[LIVMFYC]-[PG]-T-[DE]-[STA]-x-[FY]-[GA]- [LIVM]-[GS]-

[1] Bairoch A., Rudd K.E., Robison K. Unpublished observations (1995).

5

613. Sucrose synthase

Sucrose synthases catalyse the synthesis of sucrose from UDP-glucose and fructose. This family includes the bulk of the sucrose synthase protein. However the carboxyl terminal region of the sucrose synthases belongs to the glycosyl transferase family Glycos_transf_1.

10

614. Sulfotransferase proteins

Number of members: 59

5

615. Synaptophysin / synaptoporin signature

Synaptophysin and synaptoporin [1] are structurally related proteins, found in the membrane of synaptic vesicles, which may function as ionic or solute channels. These two glycoproteins seem to span the membrane four times. Both their N- and C-termini sequences seem to be cytoplasmically located. As a signature pattern for this family of proteins, a highly conserved region located in the beginning of the first intravesicular loop just after the first transmembrane domain has been selected. This region contains a cysteine residue that may be involved in a disulfide bond.

20

Consensus pattern: L-S-V-[DE]-C-x-N-K-T [C may be involved in a disulfide bond

25

[1] Knaus P., Marqueze-Pouey B., Scherer H., Betz H. Neuron 5:453-462(1990).

616. Syndecans signature

Syndecans [1,2] (from the greek syndein; to bind together) are a family of transmembrane heparan sulfate proteoglycans which are implicated in the binding of extracellular matrix components and growth factors. Syndecans bind a variety of molecules via their heparan sulfate chains and can act as receptors or as co-receptors. Structurally, these proteins consist

30

of four separate domains: a) A signal sequence; b) An extracellular domain (ectodomain) of variable length and whose sequence is not evolutionary conserved in the various forms of syndecans. The ectodomain contains the sites of attachment of the heparan sulfate glycosaminoglycan side chains; c) A transmembrane region; d) A highly conserved cytoplasmic domain of about 30 to 35 residues which could interact with cytoskeletal proteins. The proteins known to belong to this family are: - Syndecan 1. - Syndecan 2 or fibroglycan. - Syndecan 3 or neuroglycan or N-syndecan. - Syndecan 4 or amphiglycan or ryudocan. - Drosophila syndecan. - Caenorhabditis elegans probable syndecan (F57C7.3). The signature pattern that has been developed for syndecans starts with the last residue of the transmembrane region and includes the first 10 residues of the cytoplasmic domain. This region, which contains four basic residues, could act as a stop transfer site.

Consensus pattern: [FY]-R-[IM]-[KR]-K(2)-D-E-G-S-Y

[1] Bernfield M., Kokenyesi R., Kato M., Hinkes M.T., Spring J., Gallo R.L., Lose E.J. Annu. Rev. Cell Biol. 8:365-393(1992).[2] David G. FASEB J. 7:1023-1030(1993).

617. Syntaxin / epimorphin family signature

The following proteins have been shown to be evolutionary related [1,2,3]: - Epimorphin (or syntaxin 2), a mammalian mesenchymal protein which plays an essential role in epithelial morphogenesis. - Syntaxin 1A (also known as antigen HPC-1) and syntaxin 1B which are synaptic proteins which may be involved in docking of synaptic vesicles at presynaptic active zones. - Syntaxin 3. - Syntaxin 4, which is potentially involved in docking of synaptic vesicles at presynaptic active zones. - Syntaxin 5, which mediates endoplasmic reticulum to golgi transport. - Syntaxin 6, which is involved in intracellular vesicle trafficking. - Syntaxin 7. - Yeast PEP12 (or VPS6) which is required for the transport of proteases to the vacuole. - Yeast SED5 which is required for the fusion of transport vesicles with the Golgi complex. - Yeast SSO1 and SSO2 which are required for vesicle fusion with the plasma membrane. - Yeast VAM3, which is required for vacuolar assembly. - Arabidopsis thaliana protein KNOLLE which may be involved in cytokinesis. - Caenorhabditis elegans hypothetical proteins F35C8.4, F48F7.2, F55A11.2 and T01B11.3. The above proteins share the following characteristics: a size ranging from 30 Kd to 40 Kd; a C-terminal extremity which is highly hydrophobic and is probably involved in anchoring the protein to the membrane; a central,

513

well conserved region, which seems to be in a coiled-coil conformation. The pattern specific for this family is based on the most conserved region of the coiled coil domain.

Consensus pattern: [RQ]-x(3)-[LIVMA]-x(2)-[LIVM]-[ESH]-x(2)-[LIVMT]-x-[DEVMT]-
[LIVM]-x(2)-[LIVM]-[FS]-x(2)-[LIVM]-x(3)-[LIVT]-x(2)-Q- [GADEQ]-x(2)-[LIVM]-
5 [DNQT]-x-[LIVMF]-[DESV]-x(2)-[LIVM]

[1] Bennett M.K., Garcia-Ararras J.E., Elferink L.A., Peterson K., Fleming A.M., Hazuka C.D., Scheller R.H. Cell 74:863-873(1993).[2] Spring J., Kato M., Bernfield M. Trends Biochem. Sci. 18:124-125(1993).[3] Pelham H.R.B. Cell 73:425-426(1993).

10

618. Sm protein

The U1, U2, U4/U6, and U5 small nuclear ribonucleoprotein particles (snRNPs) involved in pre-mRNA splicing contain seven Sm proteins (B/B', D1, D2, D3, E, F and G) in common, which assemble around the Sm site present in four of the major spliceosomal small nuclear RNAs. These proteins contain a common sequence motif in two segments, Sm1 and Sm2, separated by a short variable linker.

[1] Hermann H, Fabrizio P, Raker VA, Foulaki K, Hornig H, Brahms H, Luhrmann R EMBO J 1995;14:2076-2088. [2] Kambach C, Walke S, Young R, Avis JM, de la Fortelle E, Raker VA, Luhrmann R, Li J, Nagai K; Cell 1999;96:375-387.

25 619. Skp1 family

[1] Stebbins CE, Kaelin WG Jr, Pavletich NP; Science 1999;284:455-461.

30 620. Protein secY signatures

The eubacterial secY protein [1] plays an important role in protein export. It interacts with the signal sequences of secretory proteins as well as with two other components of the protein translocation system: secA and secE. SecY is an integral plasma membrane protein of 419 to

492 amino acid residues that apparently contains ten transmembrane segments. Such a structure probably confers to secY a 'translocator' function, providing a channel for periplasmic and outer-membrane precursor proteins. Homologs of secY are found in archaebacteria [2]. SecY is also encoded in the chloroplast genome of some algae [3] where it could be involved in a prokaryotic-like protein export system across the two membranes of the chloroplast endoplasmic reticulum (CER) which is present in chromophyte and cryptophyte algae. Two signature patterns have been developed for secY proteins. The first corresponds to the second transmembrane region, which is the most conserved section of these proteins. The second spans the C-terminal part of the fourth transmembrane region, a short intracellular loop, and the N-terminal part of the fifth transmembrane region.

Consensus pattern: [GST]-[LIVMF](2)-x-[LIVM]-G-[LIVM]-x-P-[LIVMFY](2)-x-[AS]-[GSTQ]-[LIVMFAT](3)-Q-[LIVMFA](2)

Consensus pattern: [LIVMFYW](2)-x-[DE]-x-[LIVMF]-[STN]-x(2)-G-[LIVMF]-[GST]-[NST]-G-x-[GST]-[LIVMF](3)

[1] Ito K. Mol. Microbiol. 6:2423-2428(1992). [2] Auer J., Spicker G., Boeck A. Biochimie 73:683-688(1991). [3] Douglas S.E. FEBS Lett. 298:93-96(1992).

621. (Seed protein) Small hydrophilic plant seed proteins signature. The following small hydrophilic plant seed proteins are structurally related: - Arabidopsis thaliana proteins GEA1 and GEA6. - Cotton late embryogenesis abundant (LEA) protein D-19. - Carrot EMB-1 protein. - Barley LEA proteins B19.1A, B19.1B, B19.3 and B19.4. - Maize late embryogenesis abundant protein Emb564. - Radish late seed maturation protein p8B6. - Rice embryonic abundant protein Emp1. - Sunflower 10 Kd late embryogenesis abundant protein (DS10). - Wheat Em proteins. These proteins contains from 83 to 153 amino acid residues and may play a role [1,2] in equipping the seed for survival, maintaining a minimal level of hydration in the dry organism and preventing the denaturation of cytoplasmic components. They may also play a role during imbibition by controlling water uptake. As a signature pattern, the best conserved region in the sequence of these proteins has been developed, it is a glycine-rich nonapeptide located in the N-terminal section.-

Consensus pattern: G-[EQ]-T-V-V-P-G-G-T-

[1] Dure L. III, Crouch M., Harada J., Ho T.-H. D., Mundy J., Quatrano R., Thomas T., Sung Z.R. *Plant Mol. Biol.* 12:475-486(1989).[2] Gaubier P., Raynal M., Hull G., Huestis G.M., Grellet F., Arenas C., Pages M., Delseny M. *Mol. Gen. Genet.* 238:409-418(1993).

5

622. Serine carboxypeptidases, active sites

All known carboxypeptidases are either metallo carboxypeptidases or serinecarboxypeptidases. The catalytic activity of the serine carboxypeptidases, like that of the trypsin family serine proteases, is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which is itself hydrogen-bonded to a serine [1].

10

Proteins known to be serine carboxypeptidases are: - Barley and wheat serine carboxypeptidases I, II, and III [2]. - Yeast carboxypeptidase Y (YSCY) (gene PRC1), a vacuolar protease involved in degrading small peptides. - Yeast KEX1 protease, involved in killer toxin and alpha-factor precursor processing. - Fission yeast *sxa2*, a probable carboxypeptidase involved in degrading or processing mating pheromones [3]. - *Penicillium janthinellum* carboxypeptidase S1 [4]. - *Aspergillus niger* carboxypeptidase pepF. - *Aspergillus sato*i carboxypeptidase cpdS. - Vertebrate protective protein / cathepsin A [5], a lysosomal protein which is not only a carboxypeptidase but also essential for the activity of both beta-galactosidase and neuraminidase. - Mosquito vitellogenic carboxypeptidase (VCP) [6]. - *Naegleria fowleri* virulence-related protein Nf314 [7]. - Yeast hypothetical protein YBR139w. - *Caenorhabditis elegans* hypothetical proteins C08H9.1, F13D12.6, F32A5.3, F41C3.5 and K10B2.2. This family also includes: - Sorghum (s)-hydroxymandelonitrile lyase (hydroxynitrile lyase) (HNL) [8], an enzyme involved in plant cyanogenesis. The sequences surrounding the active site serine and histidine residues are highly conserved in all these

25

serine carboxypeptidases.

Consensus pattern: [LIVM]-x-[GTA]-E-S-Y-[AG]-[GS] [S is the active site residue]

Consensus pattern: [LIVF]-x(2)-[LIVSTA]-x-[IVPST]-x-[GSDNQL]-[SAGV]-[SG]-H-x-[IVAQ]-P-x(3)-[PSA] [H is the active site residue]

30

[1] Liao D.I., Remington S.J. *J. Biol. Chem.* 265:6528-6531(1990).[2] Sorensen S.B., Svendsen I., Breddam K. *Carlsberg Res. Commun.* 54:193-202(1989).[3] Imai Y., Yamamoto M. *Mol. Cell. Biol.* 12:1827-1834(1992).[4] Svendsen I., Hofmann T., Endrizzi J., Remington J., Breddam K. *FEBS Lett.* 333:39-43(1993).[5] Galjart N.J., Morreau H., Willemsen R., Gillemans N., Bonten E.J., d'Azzo A. *J. Biol. Chem.* 266:14754-14762(1991).[

- 6] Cho W.L., Deitsch K.W., Raikhel A.S. Proc. Natl. Acad. Sci. U.S.A. 88:10821-10824(1991).[7] Hu W.N., Kopachik W., Band R.N. Infect. Immun. 60:2418-2424(1992).[8] Wajant H., Mundry K.W., Pfitzenmaier K. Plant Mol. Biol. 26:735-746(1994).[9] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

5

623. Serpins signature. Serpins (SERine Proteinase INhibitors) [1,2,3,4] are a group of structurally related proteins. They are high molecular weight (400 to 500 amino acids), extracellular, irreversible serine protease inhibitors with a well defined structural-functional characteristic: a reactive region that acts as a 'bait' for an appropriate serine protease. This region is found in the C-terminal part of these proteins. Proteins which are known to belong to the serpin family are listed below (references are only provided for recently determined sequences): - Alpha-1 protease inhibitor (alpha-1-antitrypsin, contrapsin). - Alpha-1-antichymotrypsin, - Antithrombin III. - Alpha-2-antiplasmin. - Heparin cofactor II. - Complement C1 inhibitor. - Plasminogen activator inhibitors 1 (PAI-1) and 2 (PAI-2). - Glia derived nexin (GDN) (Protease nexin I). - Protein C inhibitor. - Rat hepatocytes SPI-1, SPI-2 and SPI-3 inhibitors. - Human squamous cell carcinoma antigen (SCCA) which may act in the modulation of the host immune response against tumor cells. - A lepidopteran protease inhibitor. - Leukocyte elastase inhibitor which, in contrast to other serpins, is an intracellular protein. - Neuroserpin [5], a neuronal inhibitor of plasminogen activators and plasmin. - Cowpox virus crmA [6], an inhibitor of the thiol protease interleukin-1B converting enzyme (ICE). CrmA is the only serpin known to inhibit a non-serine proteinase. - Some orthopoxviruses probable protease inhibitors, which may be involved in the regulation of the blood clotting cascade and/or of the complement cascade in the mammalian host. On the basis of strong sequence similarities, a number of proteins with no known inhibitory activity are said to belong to this family: - Birds ovalbumin and the related genes X and Y proteins. - Angiotensinogen; the precursor of the angiotensin active peptide. - Barley protein Z; the major endosperm albumin. - Corticosteroid binding globulin (CBG). - Thyroxine-binding globulin (TBG). - Sheep uterine milk protein (UTMP) and pig uteroferrin-associated protein (UFAP). - Hsp47, an endoplasmic reticulum heat-shock protein that binds strongly to collagen and could act as a chaperone in the collagen biosynthetic pathway [7]. - Maspin, which seems to function as a tumor suppressor [5]. - Pigment epithelium-derived factor precursor (PEDF), a protein with a strong neutrophilic activity [8]. -

10

623. Serpins signature

5

20

25

30

Ep45, an estrogen-regulated protein from *Xenopus* [9]. A signature pattern has been developed for this family of proteins, centered on a well conserved Pro-Phe sequence which is found ten to fifteen residues on the C-terminal side of the reactive bond

- 5 Consensus pattern: [LIVMFY]-x-[LIVMFYAC]-[DNQ]-[RKHQS]-[PST]-F-[LIVMFY]-[LIVMFYC]-x-[LIVMFAH]-

[1] Carrell R., Travis J. Trends Biochem. Sci. 10:20-24(1985).[2] Carrell R., Pemberton P.A., Boswell D.R. Cold Spring Harbor Symp. Quant. Biol. 52:527-535(1987).[3] Huber R., Carrell R.W. Biochemistry 28:8951-8966(1989).[4] Remold-O'Donneel E. FEBS Lett. 315:105-108(1993).[5] Osterwalder T., Contartese J., Stoeckli E.T., Kuhn T.B., Sonderegger P. EMBO J. 15:2944-2953(1996).[6] Komiyama T., Ray C.A., Pickup D.J., Howard A.D., Thornberry N.A., Peterson E.P., Salvesen G. J. Biol. Chem. 269:19331-19337(1994).[7] Clarke E., Sandwal B.D. Biochim. Biophys. Acta 1129:246-248(1992).[8] Zou Z., Anisowicz A., Neveu M., Rafidi K., Sheng S., Sager R., Hendrix M.J., Seftor E., Thor A. Science 263:526-529(1994).[9] Steele F.R., Chader G.J., Johnson L.V., Tombran-Tink J. Proc. Natl. Acad. Sci. U.S.A. 90:1526-1530(1993).[10] Holland L.J., Suksang C., Wall A.A., Roberts L.R., Moser D.R., Bhattacharya A. J. Biol. Chem. 267:7053-7059(1992).

624. Sigma-54 interaction domain signatures and profile

Some bacterial regulatory proteins activate the expression of genes from promoters recognized by core RNA polymerase associated with the alternative sigma-54 factor. These have a conserved domain of about 230 residues involved in the ATP-dependent [1,2] interaction with sigma-54. This domain has been found in the proteins listed below: - acoR from *Alcaligenes eutrophus*, an activator of the acetoin catabolism operon acoXABC. - algB from *Pseudomonas aeruginosa*, an activator of alginate biosynthetic gene algD. - dctD from *Rhizobium*, an activator of dctA, the C4-dicarboxylate transport protein. - dhaR from *Citrobacter freundii*, a regulator of the dha operon for glycerol utilization. - fhIA from *Escherichia coli*, an activator of the formate dehydrogenase H and hydrogenase III structural genes. - flbD from *Caulobacter crescentus*, an activator of flagellar genes. - hoxA from *Alcaligenes eutrophus*, an activator of the hydrogenase operon. - hrpS from *Pseudomonas syringae*, an activator of hprD as well as other hrp loci involved in plant pathogenicity. -

hupR1 from *Rhodobacter capsulatus*, an activator of the [NiFe] hydrogenase genes hupSL. - hydG from *Escherichia coli* and *Salmonella typhimurium*, an activator of the hydrogenase activity. - levR from *Bacillus subtilis*, which regulates the expression of the levanase operon (levDEFG and sacC). - nifA (as well as anfA and vnfA) from various bacteria, an activator of the nif nitrogen-fixing operon. - ntrC, from various bacteria, an activator of nitrogen assimilatory genes such as that for glutamine synthetase (glnA) or of the nif operon. - pgdA from *Salmonella typhimurium*, the activator of the inducible phospho- glycerate transport system. - pilR from *Pseudomonas aeruginosa*, an activator of pilin gene transcription. - rocR from *Bacillus subtilis*, an activator of genes for arginine utilization - tyrR from *Escherichia coli*, involved in the transcriptional regulation of aromatic amino-acid biosynthesis and transport. - wtsA, from *Erwinia stewartii*, an activator of plant pathogenicity gene wtsB. - xylR from *Pseudomonas putida*, the activator of the tol plasmid xylene catabolism operon xylCAB and of xylS. - *Escherichia coli* hypothetical protein yfhA. - *Escherichia coli* hypothetical protein yhgB. About half of these proteins (algB, dcdT, flbD, hoxA, hupR1, hydG, ntrC, pgdA and pilR) belong to signal transduction two-component systems [3] and possess a domain that can be phosphorylated by a sensor-kinase protein in their N- terminal section. Almost all of these proteins possess a helix-turn-helix DNA-binding domain in their C-terminal section. The domain which interacts with the sigma-54 factor has an ATPase activity. This may be required to promote a conformational change necessary for the interaction [4]. The domain contains an atypical ATP-binding motif A (P-loop) as well as a form of motif B. The two ATP-binding motifs are located in the N-terminal section of the domain; signature patterns have been developed for both motifs. Other regions of the domain are also conserved. One of them, located in the C-terminal section, has been selected as a third signature pattern.

Consensus pattern: [LIVMFY](3)-x-G-[DEQ]-[STE]-G-[STAV]-G-K-x(2)-[LIVMFY]

Consensus pattern: [GS]-x-[LIVMF]-x(2)-A-[DNEQASH]-[GNEK]-G-[STIM]-

[LIVMFY](3)-[DE]-[EK]-[LIVM]

Consensus pattern: [FYW]-P-[GS]-N-[LIVM]-R-[EQ]-L-x-[NHAT]

[1] Morrett E., Segovia L. J. *Bacteriol.* 175:6067-6074(1993).[2] Austin S., Kundrot C.,

Dixon R. *Nucleic Acids Res.* 19:2281-2287(1991).[3] Albright L.M., Huala E., Ausubel

F.M. *Annu. Rev. Genet.* 23:311-336(1989).[4] Austin S., Dixon R. *EMBO J.* 11:2219-

2228(1992).

625. Sigma-70 factors family signatures

Sigma factors [1] are bacterial transcription initiation factors that promote the attachment of the core RNA polymerase to specific initiation sites and are then released. They alter the specificity of promoter recognition. Most bacteria express a multiplicity of sigma factors. Two of these factors, sigma-70 (gene *rpoD*), generally known as the major or primary sigma factor, and sigma-54 (gene *rpoN* or *ntrA*) direct the transcription of a wide variety of genes. The other sigma factors, known as alternative sigma factors, are required for the transcription of specific subsets of genes. With regard to sequence similarity, sigma factors can be grouped into two classes: the sigma-54 and sigma-70 families. The sigma-70 family includes, in addition to the primary sigma factor, a wide variety of sigma factors, some of which are listed below: - *Bacillus* sigma factors involved in the control of sporulation-specific genes: sigma-E (sigE or *spoIIGB*), sigma-F (sigF or *spoIIAC*), sigma-G (sigG or *spoIIIG*), sigma-H (sigH or *spo0C*) and sigma-K (sigK or *spoIVCB/spoIIIC*). - *Escherichia coli* and related bacteria sigma-32 (gene *rpoH* or *htpR*) involved in the expression of heat shock genes. - *Escherichia coli* and related bacteria sigma-27 (gene *fliA*) involved in the expression of the flagellin gene. - *Escherichia coli* sigma-S (gene *rpoS* or *katF*) which seems to be involved in the expression of genes required for protection against external stresses. - *Myxococcus xanthus* sigma-B (sigB) which is essential for the late-stage differentiation of that bacteria. Alignments of the sigma-70 family permit the identification of four regions of high conservation [2,3]. Each of these four regions can in turn be subdivided into a number of sub-regions. Signature patterns based on the two best-conserved sub-regions have been developed. The first pattern corresponds to sub-region 2.2; the exact function of this sub-region is not known although it could be involved in the binding of the sigma factor to the core RNA polymerase. The second pattern corresponds to sub-region 4.2 which seems to harbor a DNA-binding 'helix-turn-helix' motif involved in binding the conserved -35 region of promoters recognized by the major sigma factors. The second pattern starts one residue before the N-terminal extremity of the HTH region and ends six residues after its C-terminal extremity.

Consensus pattern: [DE]-[LIVMF](2)-[HEQS]-x-G-x-[LIVMFA]-G-L-[LIVMFYE]-x-[GSAM]-[LIVMAP]

Consensus pattern: [STN]-x(2)-[DEQ]-[LIVM]-[GAS]-x(4)-[LIVMF]-[PSTG]-x(3)-[LIVMA]-x-[NQR]-[LIVMA]-[EQH]-x(3)-[LIVMFW]-x(2)-[LIVM]

[1] Helmann J.D., Chamberlin M.J. Annu. Rev. Biochem. 57:839-872(1988).[2] Gribskov M., Burgess R.R. Nucleic Acids Res. 14:6745-6763(1986).[3] Lonetto M.A., Gribskov M., Gross C.A. J. Bacteriol. 174:3843-3849(1992).[4] Lonetto M.A., Brown K.L., Rudd K.E., Buttner M.J. Proc. Natl. Acad. Sci. U.S.A. 91:7573-7577(1994).

5

626. Signal carboxyl-terminal domain. 430 members.

10 627. Signal peptidases I signatures

Signal peptidases (SPases) [1] (also known as leader peptidases) remove the signal peptides from secretory proteins. In prokaryotes three types of SPases are known: type I (gene *lepB*) which is responsible for the processing of the majority of exported pre-proteins; type II (gene *lsp*) which only process lipoproteins, and a third type involved in the processing of pili subunits. SPase I is an integral membrane protein that is anchored in the cytoplasmic membrane by one (in *B. subtilis*) or two (in *E. coli*) N-terminal transmembrane domains with the main part of the protein protruding in the periplasmic space. Two residues have been shown [2,3] to be essential for the catalytic activity of SPase I: a serine and an lysine. SPase I is evolutionary related to the yeast mitochondrial inner membrane protease subunit 1 and 2 (genes *IMP1* and *IMP2*) which catalyze the removal of signal peptides required for the targeting of proteins from the mitochondrial matrix, across the inner membrane, into the inter-membrane space [4]. In eukaryotes the removal of signal peptides is effected by an oligomeric enzymatic complex composed of at least five subunits: the signal peptidase complex (SPC). The SPC is located in the endoplasmic reticulum membrane. Two components of mammalian SPC, the 18 Kd (SPC18) and the 21 Kd (SPC21) subunits as well as the yeast SEC11 subunit have been shown [5] to share regions of sequence similarity with prokaryotic SPases I and yeast *IMP1/IMP2*. Three signature patterns for these proteins have been developed. The first signature contains the putative active site serine, the second signature contains the putative active site lysine which is not conserved in the SPC subunits, and the third signature corresponds to a conserved region of unknown biological significance which is located in the C-terminal section of all these proteins.

Consensus pattern: [GS]-x-S-M-x-[PS]-[AT]-[LF] [S is an active site residue]

Consensus pattern: K-R-[LIVMSTA](2)-G-x-[PG]-G-[DE]-x-[LIVM]-x-[LIVMFY] [K is an active site residue]

Consensus pattern: [LIVMFYW](2)-x(2)-G-D-[NH]-x(3)-[SND]-x(2)-[SG]

[1] Dalbey R.E., von Heijne G. Trends Biochem. Sci. 17:474-478(1992).[2] Sung M.,

- 5 Dalbey R.E. J. Biol. Chem. 267:13154-13159(1992).[3] Black M.T. J. Bacteriol. 175:4957-4961(1993).[4] Nunnari J., Fox T.D., Walter P. Science 262:1997-2004(1993).[5] van Dijk J.M., de Jong A., Vehmaanpera J., Venema G., Bron S. EMBO J. 11:2819-2828(1992).[6] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

10

628. (sodcu) Copper/Zinc superoxide dismutase signatures

Copper/Zinc superoxide dismutase (SODC) [1] is one of the three forms of an enzyme that catalyzes the dismutation of superoxide radicals. SODC binds one atom each of zinc and copper. Various forms of SODC are known: acytoplasmic form in eukaryotes, an additional chloroplast form in plants, an extracellular form in some eukaryotes, and a periplasmic form in prokaryotes. The metal binding sites are conserved in all the known SODC sequences [2]. Two signature patterns have been derived for this family of enzymes: the first one contains two histidine residues that bind the copper atom; the second one is located in the C-terminal section of SODC and contains a cysteine which is involved in a disulfide bond.

Consensus pattern: [GA]-[IMFAT]-H-[LIVF]-H-x(2)-[GP]-[SDG]-x-[STAGDE] [The two H's are copper ligands]

Consensus pattern: G-[GN]-[SGA]-G-x-R-x-[SGA]-C-x(2)-[IV] [C is involved in a disulfide bond]

[1] Bannister J.V., Bannister W.H., Rotilio G. CRC Crit. Rev. Biochem. 22:111-154(1987).[

- 25 2] Smith M.W., Doolittle R.F. J. Mol. Evol. 34:175-184(1992).

629. (sodfe) Manganese and iron superoxide dismutases signature

Manganese superoxide dismutase (SODM) [1] is one of the three forms of an enzyme that catalyzes the dismutation of superoxide radicals. The four ligands of the manganese atom are conserved in all the known SODM sequences. These metal ligands are also conserved in the related iron form of superoxide dismutases [2,3]. A short conserved region which includes two of the four ligands: an aspartate and a histidine has been selected as a signature.

30

522

Consensus pattern: D-x-W-E-H-[STA]-[FY](2) [D and H are manganese/iron ligands]

[1] Bannister J.V., Bannister W.H., Rotilio G. CRC Crit. Rev. Biochem. 22:111-154(1987).
[2] Parker M.W., Blake C.C.F. FEBS Lett. 229:377-382(1988). [3] Smith M.W., Doolittle
R.F. J. Mol. Evol. 34:175-184(1992).

5

630. Spectrin repeat

Spectrin repeats are found in several proteins involved in cytoskeletal structure. These include spectrin, alpha-actinin and dystrophin. The sequence repeat used in this family is taken from the structural repeat in reference [2]. The spectrin repeat forms a three helix bundle. The second helix is interrupted by proline in some sequences.

Number of members: 898

[1] Actin-binding proteins. 1: Spectrin super family. Hartwig JH; Protein Profile 1995;2:732-732. [2] Crystal structure of the repetitive segments of spectrin. Yan Y, Winograd E, Viel A, Cronin T, Harrison SC, Branton D; Science 1993;262:2027-2030.

631. (subtilase) Streptomyces subtilisin-type inhibitors signature

Bacteria of the Streptomyces family produce a family of proteinase inhibitors[1] characterized by their strong activity toward subtilisin. They are collectively known as SSI's: Streptomyces Subtilisin Inhibitors. Some SSI's also inhibit trypsin or chymotrypsin. In their mature secreted form, SSI's are proteins of about 110 residues with two conserved disulfide bonds. +-----+ +-----+ |||

xxxxxxxxxxxxxxxxCxxxxxxxxCxxxxxxxxCx#xxxxxxxxxxxxxxxxCxxxxxx *****'C':
conserved cysteine involved in a disulfide bond. '#': active site residue. '*': position of the pattern.

Consensus pattern: C-x-P-x(2,3)-G-x-H-P-x(4)-A-C-[ATD]-x-L [The two C's are involved in a disulfide bond]

[1] Taguchi S., Kojima S., Terabe M., Miura K.-I., Momose H. Eur. J. Biochem. 220:911-918(1994).

632. Sugar transport proteins signatures

In mammalian cells the uptake of glucose is mediated by a family of closely related transport proteins which are called the glucose transporters [1,2,3]. At least seven of these transporters are currently known to exist (in Human they are encoded by the GLUT1 to GLUT7 genes). These integral membrane proteins are predicted to comprise twelve membrane spanning domains. The glucose transporters show sequence similarities [4,5] with a number of other sugar or metabolite transport proteins listed below (references are only provided for recently determined sequences). - *Escherichia coli* arabinose-proton symport (araE). - *Escherichia coli* galactose-proton symport (galP). - *Escherichia coli* and *Klebsiella pneumoniae* citrate-proton symport (also known as citrate utilization determinant) (gene cit). - *Escherichia coli* alpha-ketoglutarate permease (gene kgtP). - *Escherichia coli* proline/betaine transporter (gene proP) [6]. - *Escherichia coli* xylose-proton symport (xylE). - *Zymomonas mobilis* glucose facilitated diffusion protein (gene glf). - Yeast high and low affinity glucose transport proteins (genes SNF3, HXT1 to HXT14). - Yeast galactose transporter (gene GAL2). - Yeast maltose permeases (genes MAL3T and MAL6T). - Yeast myo-inositol transporters (genes ITR1 and ITR2). - Yeast carboxylic acid transporter protein homolog JEN1. - Yeast inorganic phosphate transporter (gene PHO84). - *Kluyveromyces lactis* lactose permease (gene LAC12). - *Neurospora crassa* quinate transporter (gene Qa-y), and *Emericella nidulans* quinate permease (gene qutD). - *Chlorella hexose* carrier (gene HUP1). - *Arabidopsis thaliana* glucose transporter (gene STP1). - Spinach sucrose transporter. - *Leishmania donovani* transporters D1 and D2. - *Leishmania enriettii* probable transport protein (LTP). - Yeast hypothetical proteins YBR241c, YCR98c and YFL040w. - *Caenorhabditis elegans* hypothetical protein ZK637.1. - *Escherichia coli* hypothetical proteins yabE, ydjE and yhjE. - *Haemophilus influenzae* hypothetical proteins HI0281 and HI0418. - *Bacillus subtilis* hypothetical proteins yxbC and yxdF. It has been suggested [4] that these transport proteins have evolved from the duplication of an ancestral protein with six transmembrane regions, this hypothesis is based on the conservation of two G-R-[KR] motifs. The first one is located between the second and third transmembrane domains and the second one between transmembrane domains 8 and 9. Two patterns have been developed to detect this family of proteins. The first pattern is based on the G-R-[KR] motif; but because this motif is too short to be specific to this family of proteins, a pattern from a larger region centered on the second copy of this motif was derived. The second pattern is based on a

number of conserved residues which are located at the end of the fourth transmembrane segment and in the short loop region between the fourth and fifth segments.

Consensus pattern: [LIVMSTAG]-[LIVMFSAG]-x(2)-[LIVMSA]-[DE]-x-[LIVMFYWA]-G- R-[RK]-x(4,6)-[GSTA]

5 Consensus pattern: [LIVMF]-x-G-[LIVMFA]-x(2)-G-x(8)-[LIFY]-x(2)-[EQ]-x(6)- [RK]
 [1] Silverman M. Annu. Rev. Biochem. 60:757-794(1991).[2] Gould G.W., Bell G.I. Trends Biochem. Sci. 15:18-23(1990).[3] Baldwin S.A. Biochim. Biophys. Acta 1154:17-49(1993).[4] Maiden M.C.J., Davis E.O., Baldwin S.A., Moore D.C.M., Henderson P.J.F. Nature 325:641-643(1987).[5] Henderson P.J.F. Curr. Opin. Struct. Biol. 1:590-601(1991).[6]
 10 Culham D.E., Lasby B., Marangoni A.G., Milner J.L., Steer B.A., van Nues R.W., Wood J.M. J. Mol. Biol. 229:268-276(1993).

633. Synaptobrevin signature

15 Synaptobrevin [1] is an intrinsic membrane protein of small synaptic vesicles whose function is not yet known, but which is highly conserved in mammals, electric ray (where its is known as VAMP-1), Drosophila and yeast [2]. In yeast there are two closely related forms of synaptobrevin (genes SNC1 and SNC2) while in mammals there is at least 4 (genes SYB1, SYB2, SYB3 and SYBL1). Structurally synaptobrevin consist of a N-terminal cytoplasmic
 20 domain of from 90 to 110 residues, followed by a transmembrane region, and then by a short (from 2 to 22 residues) C-terminal intravesicular domain. As a signature pattern for synaptobrevin, a highly conserved stretch of residues located in the central part of the sequence was selected.

Consensus pattern: N-[LIVM]-[DENS]-[KL]-V-x-[DEQ]-R-x(2)-[KR]-[LIVM]-[STDE]- x-
 25 [LIVM]-x-[DE]-[KR]-[TA]-[DE]

[1] Suedhof T.C., Baumert M., Perin M.S., Jahn R. Neuron 2:1475-1481(1989).[2] Gerst J.E., Rodgers L., Riggs M., Wigler M. Proc. Natl. Acad. Sci. U.S.A. 89:4338-4342(1992).

634. TBC domain. Identification of a TBC domain in GYP6_YEAST and GYP7_YEAST, which are GTPase activator proteins of yeast Ypt6 and Ypt7, imply that these domains are
 30 GTPase activator proteins of Rab-like small GTPases. Number of members: 55

[1] Medline: 96032578. Molecular cloning of a cDNA with a novel domain present in the tre-2 oncogene and the yeast cell cycle regulators BUB2 and cdc16. Richardson PM, Zon LI; Oncogene 1995;11:1139-1148.

[2]Medline: 97398935. A shared domain between a spindle assembly checkpoint protein and Ypt/Rab-specific GTPase-activators. Neuwald AF; Trends Biochem Sci 1997;22:243-244.

635. Transcription factor TFIID repeat signature (TBP)

Transcription factor TFIID (or TATA-binding protein, TBP) [1,2] is a general factor that plays a major role in the activation of eukaryotic genes transcribed by RNA polymerase II. TFIID binds specifically to the TATA box promoter element which lies close to the position of transcription initiation. There is a remarkable degree of sequence conservation of a C-terminal domain of about 180 residues in TFIID from various eukaryotic sources. This region is necessary and sufficient for TATA box binding. The most significant structural feature of this domain is the presence of two conserved repeats of a 77 amino-acid region. The intramolecular symmetry generates a saddle-shaped structure that sits astride the DNA [3]. Drosophila TRF (TBP-related factor) [4] is a sequence-specific transcription factor that also binds to the TATA box and is highly similar to TFIID. Archaeobacteria also possess a TBP homolog [5]. A signature pattern that spans the last 50 residues of the repeated region has been derived.-

Consensus pattern: Y-x-P-x(2)-[IF]-x(2)-[LIVM](2)-x-[KRH]-x(3)-P-[RKQ]-x(3)- L-[LIVM]-F-x-[STN]-G-[KR]-[LIVM]-x(3)-G-[TAGL]-[KR]-x(7)- [AGC]-x(7)-[LIVM]
[1] Hoffmann A., Sinn E., Yamamoto T., Wang J., Roy A., Horikoshi M., Roeder R.G. Nature 346:387-390(1990).[2] Gash A., Hoffmann A., Horikoshi M., Roeder R.G., Chua N.-H. Nature 346:390-394(1990).[3] Nikolov D.B., Hu S.-H., Lin J., Gasch A., Hoffmann A., Horikoshi M., Chua N.-H., Roeder R.G., Burley S.K. Nature 360:40-46(1992).[4] Crowley T.E., Hoey T., Liu J.-K., Jan Y.N., Jan L.Y., Tjian R. Nature 361:557-561(1993).[5] Marsh T.L., Reich C.I., Whitelock R.B., Olsen G.J. Proc. Natl. Acad. Sci. U.S.A. 91:4180-4184(1994).

636. Translationally controlled tumor protein signatures (TCTP)

Mammalian translationally controlled tumor protein (TCTP) (or P23) is a protein which has been found to be preferentially synthesized in cells during the early growth phase of some types of tumor [1,2], but which is also expressed in normal cells. The physiological function of TCTP is still not known. It is a hydrophilic protein of 18 to 20 Kd. Close homologs have been found in plants [3], earthworm [4], *Caenorhabditis elegans* (F52H2.11), *Hydra*, budding yeast (YKL056c) [5] and fission yeast (SpAC1F12.02c). Two of the best conserved regions have been selected as signature patterns for TCTP.

Consensus pattern: [IFA]-[GA]-[GAS]-N-[PAK]-S-[GA]-E-[GDE]-[PAGE]-[DEQGA]

Consensus pattern: [FLVH]-[FY]-[IVCT]-G-E-x-[MA]-x(2,5)-[DEN]-[GAST]-x-[LV]-

[AV]-x(3)-[FYW]

[1] Boehm H., Beendorf R., Gaestel M., Gross B., Nuernberg P., Kraft R., Otto A., Bielka H. *Biochem. Int.* 19:277-286(1989). [2] Makrides S., Chitpatima S.T., Bandyopadhyay R., Brawerman G. *Nucleic Acids Res.* 16:2350-2350(1988). [3] Pay A., Heberle-Bors E., Hirt H. *Plant Mol. Biol.* 19:501-503(1992). [4] Stuerzenbaum S.R., Kille P., Morgan A.J. *Biochim. Biophys. Acta* 1398:294-304(1998). [5] Rasmussen S.W. *Yeast* 10:S63-S68(1994).

637. TFIIS zinc ribbon domain signature

Transcription factor S-II (TFIIS) [1] is a eukaryotic protein necessary for efficient RNA polymerase II transcription elongation, past template-encoded pause sites. TFIIS shows DNA-binding activity only in the presence of RNA polymerase II. It is a protein of about 300 amino acids whose sequence is highly conserved in mammals, *Drosophila*, yeast (where it was first known as PPR2, a transcriptional regulator of *URA4*, and then as DST1, the DNA strand transfer protein alpha [2]) and in the archaeobacteria *Sulfolobus acidocaldarius* [3]. This family also includes the eukaryotic and archaeobacterial RNA polymerase subunits of the 15 Kd / M family (see <PDOC00790>) as well as the following viral proteins: - Vaccinia virus RNA polymerase 30 Kd subunit (rpo30) [4]. - African swine fever virus protein I243L [5]. The best conserved region of all these proteins contains four cysteines that bind a zinc ion and fold in a conformation termed a 'zinc ribbon' [6]. Besides these cysteines, there are a number of other conserved residues which can be used to help define a specific pattern for this type of domain.

Consensus pattern: C-x(2)-C-x(9)-[LIVMQSAR]-[QH]-[STQL]-[RA]-[SACR]-x-[DE]-[DET]-[PGSEA]-x(6)-C-x(2,5)-C-x(3)-[FW] [The four C's are zinc ligands]

- [1] Hirashima S., Hirai H., Nakanishi Y., Natori S. J. Biol. Chem. 263:3858-3863(1988).[2] Kipling D., Kearsey S.E. Nature 353:509-509(1991).[3] Langer D., Zillig W. Nucleic Acids Res. 21:2251-2251(1993).[4] Ahn B.-Y., Gershon P.D., Jones E.V., Moss B. Mol. Cell. Biol. 10:5433-5441(1990).[5] Rodriguez J.M., Salas M.L., Vinuela E. Virology 186:40-52(1992).[6] Qian X., Jeon C., Yoon H., Agarwal K., Weiss M.A. Nature 365:277-279(1993).

638. Tetrahydrofolate dehydrogenase/cyclohydrolase signatures (THF DHG CYH)

Enzymes that participate in the transfer of one-carbon units are involved in various biosynthetic pathways. In many of these processes the transfers of one-carbon units are mediated by the coenzyme tetrahydrofolate (THF). Various reactions generate one-carbon derivatives of THF which can be interconverted between different oxidation states by formyltetrahydrofolate synthetase(EC 6.3.4.3), methylenetetrahydrofolate dehydrogenase (EC 1.5.1.5 or EC 1.5.1.15) and methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9).The dehydrogenase and cyclohydrolase activities are expressed by a variety of multifunctional enzymes: - Eukaryotic C-1-tetrahydrofolate synthase (C1-THF synthase), which catalyzes all three reactions described above. Two forms of C1-THF synthases are known [1], one is located in the mitochondrial matrix, while the second one is cytoplasmic. In both forms the dehydrogenase/cyclohydrolase domain is located in the N-terminal section of the 900 amino acids protein and consists of about 300 amino acid residues. The C1-THF synthases are NADP- dependent. - Eukaryotic mitochondrial bifunctional dehydrogenase/cyclohydrolase [2]. This is an homodimeric NAD-dependent enzyme of about 300 amino acid residues. - Bacterial fold [3]. Fold is an homodimeric bifunctional NADP-dependent enzyme of about 290 amino acid residues. The sequence of the dehydrogenase/cyclohydrolase domain is highly conserved in all forms of the enzyme. Two conserved regions have been selected as signature patterns. The first one is located in the N-terminal part of these enzymes and contains three acidic residues. The second pattern is a highly conserved sequence of 9 amino acids which is located in the C-terminal section.

Consensus pattern: [EQ]-x-[EQK]-[LIVM](2)-x(2)-[LIVM]-x(2)-[LIVMY]-N-x-[DN]- x(5)-[LIVMF](3)-Q-L-P-[LV]

Consensus pattern: P-G-G-V-G-P-[MF]-T-[IV]

[1] Shannon K.W., Rabinowitz J.C. J. Biol. Chem. 263:7717-7725(1988).[2] Belanger C., Mackenzie R.E. J. Biol. Chem. 264:4837-4843(1989).[3] d'Ari L., Rabinowitz J.C. J. Biol. Chem. 266:23953-23958(1991).

5

639. Triosephosphate isomerase active site (TIM)

Triosephosphate isomerase (EC 5.3.1.1) (TIM) [1] is the glycolytic enzyme that catalyzes the reversible interconversion of glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. TIM plays an important role in several metabolic pathways and is essential for efficient energy production. It is a dimer of identical subunits, each of which is made up of about 250 amino-acid residues. A glutamic acid residue is involved in the catalytic mechanism [2]. The sequence around the active site residue is perfectly conserved in all known TIM's and can be used as a signature pattern for this type of enzyme.

Consensus pattern: [AV]-Y-E-P-[LIVM]-W-[SA]-I-G-T-[GK] [E is the active site residue]

[1] Lolis E., Alber T., Davenport R.C., Rose D., Hartman F.C., Petsko G.A. Biochemistry 29:6609-6618(1990).[2] Knowles J.R. Nature 350:121-124(1991).

640. Thymidine kinase cellular-type signature (TK)

Thymidine kinase (TK) (EC 2.7.1.21) is an ubiquitous enzyme that catalyzes the ATP-dependent phosphorylation of thymidine. A comparison of TK sequences has shown [1,2,3] that there are two different families of TK. One family groups together TK from herpes viruses as well as cellular thymidylate kinases, while the second family currently consists of TK from the following sources: - Vertebrates. - Bacterial. - Bacteriophage T4. - Pox viruses. - African swine fever virus (ASF). - Fish lymphocystis disease virus (FLDV). A conserved region which is located in the C-terminal section of these enzymes has been selected as a signature pattern for this family of TKA.

Consensus pattern: [GA]-x(1,2)-[DE]-x-Y-x-[STAP]-x-C-[NKR]-x-[CH]-[LIVMFYWH]

[1] Boyle D.B., Coupar B.E.H., Gibbs A.J., Seigman L.J., Both G.W. Virology 156:355-365(1987).[2] Blasco R., Lopez-Otin C., Munoz M., Bockamp E.-O., Simon-Mateo C., Vinuela E. Virology 178:301-304(1990).[3] Robertson G.R., Whalley J.M. Nucleic Acids Res. 16:11303-11317(1988).

641. Thymidine kinase from herpesvirus (TK herpes)

[1]

Medline: 96003730

- 5 Crystal structures of the thymidine kinase from herpes simplex virus type-1 in complex with deoxythymidine and ganciclovir.

Brown DG, Visse R, Sandhu G, Davies A, Rizkallah PJ, Melitz C, Summers WC, Sanderson MR;

10 Nat Struct Biol 1995;2:876-881.

Number of members: 65

642. Nuclear transition protein 2 signatures (TP2)

15 In mammals, the second stage of spermatogenesis is characterized by the conversion of nucleosomal chromatin to the compact, non-nucleosomal and transcriptionally inactive form found in the sperm nucleus. This condensation is associated with a double-protein transition. The first transition corresponds to the replacement of histones by several spermatid-specific proteins, also called transition proteins, which are themselves replaced by protamines during the second transition. Nuclear transition protein 2 (TP2) is one of those spermatid-specific proteins. TP2 is a basic, zinc-binding protein [1] of 116 to 137 amino-acid residues.

20 Structurally, TP2 consists of three distinct parts: a conserved serine-rich N-terminal domain of about 25 residues, a variable central domain of 20 to 50 residues which contains cysteine residues, and a conserved C-terminal domain of about 70 residues rich in lysines and arginines. Two signature patterns for TP2 have been developed: one located in the N-terminal domain, the other in the C-terminal.

Consensus pattern: H-x(3)-H-S-[NS]-S-x-P-Q-S

Consensus pattern: K-x-R-K-x(2)-E-G-K-x(2)-K-[KR]-K

[1] Baskaran R., Rao M.R.S. Biochem. Biophys. Res. Commun. 179:1491-1499(1991).

643. Thiamine pyrophosphate enzymes signature (TTP enzymes)

530

A number of enzymes require thiamine pyrophosphate (TPP) (vitamin B1) as a cofactor. It has been shown [1] that some of these enzymes are structurally related. These related TPP enzymes are: - Pyruvate oxidase (POX) (EC 1.2.3.3) Reaction catalyzed: pyruvate + orthophosphate + O(2) + H(2)O = acetyl phosphate + CO(2) + H(2)O(2). - Pyruvate decarboxylase (PDC) (EC 4.1.1.1) Reaction catalyzed: pyruvate = acetaldehyde + CO(2). - Indolepyruvate decarboxylase (EC 4.1.1.74) [2] Reaction catalyzed: indole-3-pyruvate = indole-3-acetaldehyde + CO(2). - Acetolactate synthase (ALS) (EC 4.1.3.18) Reaction catalyzed: 2 pyruvate = acetolactate + CO(2). - Benzoylformate decarboxylase (BFD) (EC 4.1.1.7) [3] Reaction catalyzed: benzoylformate = benzaldehyde + CO(2). A conserved region which is located in their C-terminal section has been selected as a signature pattern for these enzymes.

Consensus pattern: [LIVMF]-[GSA]-x(5)-P-x(4)-[LIVMFYW]-x-[LIVMF]-x-G-D-[GSA]-[GSAC]

[1] Green J.B.A. FEBS Lett. 246:1-5(1989).[2] Koga J., Adachi T., Hidaka H. Mol. Gen. Genet. 226:10-16(1991).[3] Tsou A.Y., Ransom S.C., Gerlt J.A., Buechter D.D., Babbitt P.C., Kenyon G.L. Biochemistry 29:9856-9862(1990).

644. TPR Domain

[1]

Medline: 95397415

Tetratricopeptide repeat interactions: to TPR or not to TPR?

Lamb JR, Tugendreich S, Hieter P;

Trends Biochem Sci 1995;20:257-259.

[2]Medline: 98151343

The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions.

Das AK, Cohen PW, Barford D;

EMBO J 1998;17:1192-1199.

Number of members: 621

645. Uroporphyrin-III C-methyltransferase signatures (TP methylase)

Uroporphyrin-III C-methyltransferase (EC 2.1.1.107) (SUMT) [1,2] catalyzes the transfer of two methyl groups from S-adenosyl-L-methionine to the C-2 and C-7 atoms of uroporphyrinogen III to yield precorrin-2 via the intermediate formation of precorrin-1.

SUMT is the first enzyme specific to the cobalamin pathway and precorrin-2 is a common intermediate in the biosynthesis of corrinoids such as vitamin B12, siroheme and coenzyme F430. The sequences of SUMT from a variety of eubacterial and archaebacterial species are currently available. In species such as *Bacillus megaterium* (gene *cobA*), *Pseudomonas denitrificans* (*cobA*) or *Methanobacterium ivanovii* (gene *corA*) SUMT is a protein of about 25 to 30 Kd. In *Escherichia coli* and related bacteria, the *cysG* protein, which is involved in the biosynthesis of siroheme, is a multifunctional protein composed of a N-terminal domain, probably involved in transforming precorrin-2 into siroheme, and a C-terminal domain which has SUMT activity. The sequence of SUMT is related to that of a number of *P. denitrificans* and *Salmonella typhimurium* enzymes involved in the biosynthesis of cobalamin which also seem to be SAM-dependent methyltransferases [3,4]. The similarity is especially strong with two of these enzymes: *cobI/cbiL* which encodes S-adenosyl-L-methionine--precorrin-2 methyltransferase and *cobM/cbiF* whose exact function is not known. Two signature patterns have been developed for these enzymes. The first corresponds to a well conserved region in the N-terminal extremity (called region 1 in [1,3]) and the second to a less conserved region located in the central part of these proteins (this pattern spans what are called regions 2 and 3 in [1,3]).

Consensus pattern: [LIVM]-[GS]-[STAL]-G-P-G-x(3)-[LIVMFY]-[LIVM]-T-[LIVM]-[KRHQG]-[AG]

Consensus pattern: V-x(2)-[LI]-x(2)-G-D-x(3)-[FYW]-[GS]-x(8)-[LIVF]-x(5,6)-[LIVMFYWPAC]-x-[LIVMY]-x-P-G

[1] Blanche F., Robin C., Couder M., Faucher D., Cauchois L., Cameron B., Crouzet J. J. Bacteriol. 173:4637-4645(1991).[2] Robin C., Blanche F., Cauchois L., Cameron B., Couder M., Crouzet J. J. Bacteriol. 173:4893-4896(1991).[3] Crouzet J., Cameron B., Cauchois L., Rigault S., Rouyez M.-C., Blanche F., Thibaut D., Debussche L. J. Bacteriol. 172:5980-5990(1990).[4] Roth J.R., Lawrence J.G., Rubenfield M., Kieffer-Higgins S., Church G.M. J. Bacteriol. 175:3303-3316(1993).[5] Mattheakis L.C., Shen W.H., Collier R.J. Mol. Cell. Biol. 12:4026-4037(1992).

646. Tudor domain

Domain of unknown function present in several RNA-binding proteins. copies in the Drosophila Tudor protein. Slight ambiguities in the alignment. Number of members: 18

[1]Medline: 97200561 Tudor domains in proteins that interact with RNA. Ponting CP; Trends Biochem Sci 1997;22:51-52. [2]Medline: 97157029 The human EBNA-2 coactivator p100: multidomain organization and relationship to the staphylococcal nuclease fold and to the tudor protein involved in Drosophila melanogaster development. Callebaut I, Mornon JP; Biochem J 1997;321:125-132.

647. Terpene synthase family

It has been suggested that this gene family be designated tps (for terpene synthase) [1]. It has been split into six subgroups on the basis of phylogeny, called tpsa-tpsf.

tpsa includes vetispiradiene synthase Swiss:Q39979, 5-epi-aristolochene synthase, Swiss:Q40577 and (+)-delta-cadinene synthase Swiss:P93665.

tpsb includes (-)-limonene synthase, Swiss:Q40322.

tpsc includes kaurene synthase A, Swiss:O04408.

tpsd includes taxadiene synthase, Swiss:Q41594, pinene synthase, Swiss:O24475 and myrcene synthase, Swiss:O24474.

tpse includes kaurene synthase B.

tpsf includes linalool synthase.

Number of members: 51

[1]

Medline: 97413772

Monoterpene synthases from grand fir (*Abies grandis*). cDNA isolation, characterization, and functional expression of myrcene synthase, (-)-(4S)-limonene synthase, and (-)-(1S,5S)-pinene synthase.

Bohlmann J, Steele CL, Croteau R;
J Biol Chem 1997;272:21784-21792.

648. ThiF family

This family contains a repeated domain in ubiquitin
activating enzyme E1 and members of the bacterial
ThiF/MoeB/HesA family. Number of members: 87

649. Thioester dehydrase

Members of this family are involved in fatty acid biosynthesis.
Number of members: 19

[1]

Medline: 96398612

Structure of a dehydratase-isomerase from the bacterial
pathway for biosynthesis of unsaturated fatty acids: two
catalytic activities in one active site.

Leesong M, Henderson BS, Gillig JR, Schwab JM, Smith JL;

Structure 1996;4:253-264.

Database Reference: SCOP; 1mka; fa; [SCOP-USA][CATH-PDBSUM]

Database reference: PFAMB; PB058036;

650. Tub family signatures

The mouse tubby mutation is the cause of maturity-onset obesity, insulin resistance and
sensory deficits. This mutation maps to a gene, tub [1,2], which codes for a protein that
belongs to a family which currently consists of the following members: - Mammalian tub, an
hydrophilic protein of about 500 residues, which could be involved in the hypothalamic
regulation of body weight. - Human protein TULP1 [3] which may be involved in retinis
pigmentosa 14, a retinal degeneration disease. - Mouse protein p4-6 whose function is not
known. - Caenorhabditis elegans hypothetical protein F10B5.4. - Several fragmentary
sequences from plants, Drosophila and human ESTs. While the N-terminal part of these
protein is not conserved in length nor in the sequence, the C-terminal 250 residues are highly
conserved. Therefore, two regions were selected in the C-terminal part as signature patterns.

The second region is located at the C-terminal extremity and contains a penultimate cysteine residue that could be critical to the normal functioning of these proteins.

Consensus pattern: F-[KHQ]-G-R-V-[ST]-x-A-S-V-K-N-F-Q

Consensus pattern: A-F-[AG]-I-[SAC]-[LIVM]-[ST]-S-F-x-[GST]-K-x-A-C-E

[1] Kleyn P.W., Fan W., Kovats S.G., Lee J.L., Pulido J.C., Wu Y., Berkemeier L.R., Misumi D.J., Holmgren L., Charlat O., Woolf E.A., Tayber O., Brody T., Shu P., Hawkins F., Kennedy B., Baldini L., Ebeling C., Alperin G.D., Deeds J., Lakey N.D., Culpepper J., Chen H., Gluecksmann-Kuis M.A., Carlson G.A., Duyk G.M., Moore K.J. *Cell* 85:281-290(1996).[2] Noben-Trauth K., Naggert J.K., North M.A., Nishina P.M. *Nature* 380:534-538(1996).[3] North M.A., Naggert J.K., Yan Y., Noben-Trauth K., Nishina P.M. *Proc. Natl. Acad. Sci. U.S.A.* 94:3128-3133(1997).

651. Eukaryotic DNA topoisomerase I active site

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type I topoisomerases act by catalyzing the transient breakage of DNA, one strand at a time, and the subsequent rejoining of the strands. When a eukaryotic type I topoisomerase breaks a DNA backbone bond, it simultaneously forms a protein-DNA link where the hydroxyl group of a tyrosine residue is joined to a 3'-phosphate on DNA, at one end of the enzyme-severed DNA strand. In eukaryotes and pox virus topoisomerases I, there are a number of conserved residues in the region around the active site tyrosine.

Consensus pattern: [DEN]-x(6)-[GS]-[IT]-S-K-x(2)-Y-[LIVM]-x(3)-[LIVM] [Y is the active site tyrosine]

[1] Sternglanz R. *Curr. Opin. Cell Biol.* 1:533-535(1990).[2] Sharma A., Mondragon A. *Curr. Opin. Struct. Biol.* 5:39-47(1995).[3] Lynn R.M., Bjornsti M.-A., Caron P.R., Wang J.C. *Proc. Natl. Acad. Sci. U.S.A.* 86:3559-3563(1989).[4] Roca J. *Trends Biochem. Sci.* 20:156-160(1995).[E1]

652. Transaldolase signatures

Transaldolase (EC 2.2.1.2) catalyzes the reversible transfer of a three-carbon ketol unit from sedoheptulose 7-phosphate to glyceraldehyde 3-phosphate to form erythrose 4-phosphate and

fructose 6-phosphate. This enzyme, together with transketolase, provides a link between the glycolytic and pentose-phosphate pathways. Transaldolase is an enzyme of about 34 Kd whose sequence has been well conserved throughout evolution. A lysine has been implicated [1] in the catalytic mechanism of the enzyme; it acts as a nucleophilic group that attacks the carbonyl group of fructose-6-phosphate. Transaldolase is evolutionary related [2] to a bacterial protein of about 20Kd (known as talC in *Escherichia coli*), whose exact function is not yet known. Two signature patterns have been developed for these proteins. The first, located in the N-terminal section, contains a perfectly conserved pentapeptide; these cond, includes the active site lysine.

Consensus pattern: [DG]-[IVSA]-T-[ST]-N-P-[STA]-[LIVMF](2)

Consensus pattern: [LIVM]-x-[LIVM]-K-[LIVM]-[PAS]-x-[ST]-x-[DENQPAS]-G-[LIVM]-x-[AGV]-x-[QEKRT]-x-[LIVM] [K is the active site residue]

[1] Miosga T., Schaaff-Gerstenschlaeger I., Franken E., Zimmermann F.K. *Yeast* 9:1241-1249(1993).[2] Reizer J., Reizer A., Saier M.H. Jr. *Microbiology* 141:961-971(1995).

653. (Transpeptidase) Penicillin binding protein transpeptidase domain

The active site serine (residue 337 in Swiss:P14677) is conserved in all members of this family.

[1] Pares S, Mouz N, Petillot Y, Hakenbeck R, Dideberg O *Nat Struct Biol* 1996;3:284-289.

654. Trehalase signatures

Trehalase (EC 3.2.1.28) is the enzyme responsible for the degradation of the disaccharide alpha, alpha-trehalose yielding two glucose subunits [1]. It is an enzyme found in a wide variety of organisms and whose sequence has been highly conserved throughout evolution. Two of the most highly conserved regions have been selected as signature patterns. The first pattern is located in the central section, the second one is in the C-terminal region.

Consensus pattern: P-G-G-R-F-x-E-x-Y-x-W-D-x-Y

Consensus pattern: Q-W-D-x-P-x-[GA]-W-[PAS]-P

[1] Kopp M., Mueller H., Holzer H. J. Biol. Chem. 268:4766-4774(1993).[2] Henrissat B., Bairoch A. Biochem. J. 293:781-788(1993).[E1]

5 655. Trehalose-6-phosphate synthase domain

OtsA (Trehalose-6-phosphate synthase) is homologous to regions in the subunits of yeast trehalose-6-phosphate synthase/phosphate complex, [1].

[1] Kaasen I, McDougall J, Strom AR; Gene 1994;145:9-15.

10

656. Tropomyosins signature

Tropomyosins [1,2] are family of closely related proteins present in muscle and non-muscle cells. In striated muscle, tropomyosin mediate the interactions between the troponin complex and actin so as to regulate muscle contraction. The role of tropomyosin in smooth muscle and non-muscle tissues is not clear. Tropomyosin is an alpha-helical protein that forms a coiled-coil dimer. Muscle isoforms of tropomyosin are characterized by having 284 amino acid residues and a highly conserved N-terminal region, whereas non-muscle forms are generally smaller and are heterogeneous in their N-terminal region. The signature pattern for tropomyosins is based on a very conserved region in the C-terminal section of tropomyosins and which is present in both muscle and non-muscle forms.

Consensus pattern: L-K-E-A-E-x-R-A-E

[1] Smilie L.B. Trends Biochem. Sci. 4:151-155(1979).[2] McLeod A.R. BioEssays 6:208-212(1986).

25

657. Troponin

Troponin (Tn) contains three subunits, Ca²⁺ binding (TnC), inhibitory (TnI), and tropomyosin binding (TnT). this Pfam contains members of the TnT subunit.

30

Troponin is a complex of three proteins, Ca²⁺ binding (TnC), inhibitory (TnI), and tropomyosin binding (TnT).

The troponin complex regulates Ca⁺⁺ induced muscle contraction.

This family includes troponin T and troponin I. Troponin I

binds to actin and troponin T binds to tropomyosin.

Number of members: 81 [1]

Medline: 87144593

Structure of co-crystals of tropomyosin and troponin.

- 5 White SP, Cohen C, Phillips GN Jr;
Nature 1987;325:826-828. [2]

Medline: 95155315

A direct regulatory role for troponin T and a dual role for
troponin C in the Ca²⁺ regulation of muscle contraction.

- 10 Potter JD, Sheng Z, Pan BS, Zhao J;
J Biol Chem 1995;270:2557-2562.
[3]Medline: 95324796

The troponin complex and regulation of muscle contraction.

- Farah CS, Reinach FC;
5 FASEB J 1995;9:755-767.

658. (Tryp mucin) Mucin-like glycoprotein

20 This family of trypanosomal proteins resemble vertebrate mucins. The protein consists of
three regions. The N and C termini are conserved between all members of the family,
whereas the central region is not well conserved and contains a large number of threonine
residues which can be glycosylated [1].

Indirect evidence suggested that these genes might encode the core protein of parasite
25 mucins, glycoproteins that were proposed to be involved in the interaction with, and invasion
of, mammalian host cells.

[1] Di Noia JM, Sanchez DO, Frasch AC; J Biol Chem 1995;270:24146-24149.

[2] Di Noia JM, D'Orso I, Aslund L, Sanchez DO, Frasch AC; J Biol Chem 1998;273:10843-
30 10850.

659. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. A few years ago it was found [2] that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site. The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific for arginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine, tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I synthetases [4,5,6] and seem to share the same tertiary structure based on a Rossmann fold. Consensus pattern: P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-[LIVMYAC]-G-[HNTG]-[LIVMFYSTAGPC]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M., Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

660. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1b)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. A few years ago it was found [2] that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well

conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site. The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific for arginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine, tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I

synthetases [4,5,6] and seem to share the same tertiary structure based on a Rossmann fold. Consensus pattern: P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-[LIVMYAC]-G-[HNTG]-[LIVMFYSTAGPC]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M., Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

661. (tRNA-synt 1C) tRNA synthetases class I (E and Q)

Other tRNA synthetase sub-families are too dissimilar to be included.

This family includes only glutamyl and glutaminyl tRNA synthetases.

In some organisms, a single glutamyl-tRNA synthetase aminoacylates both tRNA(Glu) and tRNA(Gln).

[1] Rath VL, Silvian LF, Beijer B, Sproat BS, Steitz TA; Structure 1998;6:439-449.

662. (tRNA-synt 1d) tRNA synthetases class I (R)

Other tRNA synthetase sub-families are too dissimilar to be included.

This family includes only arginyl tRNA synthetase.

663. Aminoacyl-transfer RNA synthetases class-II signatures (tRNA synt 2)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In

prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern: [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D. BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S. Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

664. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1e)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. A few years ago it was found [2] that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well

conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site.

The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific for arginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine, tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I synthetases [4,5,6] and seem to share the same tertiary structure based on a Rossmann fold. Consensus pattern: P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-[LIVMYAC]-G-[HNTG]-[LIVMFYSTAGPC]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M., Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

665. Aminoacyl-transfer RNA synthetases class-II signatures (tRNA synt 2b)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern: [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D. BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel

G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S.,
Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S.
Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar
N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet
5 S. Nucleic Acids Res. 18:305-312(1990).

666. Thaumatin family signature

Thaumatococcus daniellii, an African brush. The protein is made of about
10 200 residues and contains 8 disulfide bonds. A number of proteins have been found to be
related to thaumatins. These protein are listed below (references are only provided for
recently determined sequences). - A maize alpha-amylase/trypsin inhibitor. - Two tobacco
pathogenesis-related proteins: PR-R major and minor forms, which are induced after
15 infection with viruses. - Salt-induced protein NP24 from tomato. - Osmotin, a salt-induced
protein from tobacco. - Osmotin-like proteins OSML13, OSML15 and OSML81 from potato
[2]. - P21, a leaf protein from soybean. - PWIR2, a leaf protein from wheat. - Zeamatin, a
maize antifungal protein [3]. The exact biological function of all these proteins is not yet
known. A conserved region that includes three cysteine residues known (in thaumatocin) to be
20 involved in disulfide bonds has been selected as a signature pattern.

```
+-----+ | +-----+ | | ***** |
||
```

```
xxCxxxxxxxxxxxxxxxxCxxCxxCxxxxxxxxxxxxxxxxCxxCxCxxxCxCxxCCxCxxxCxxxxxC
```

```
xxxCx ||||| || | +--+ +--+ | +--+ +--+ +--+ | +-----+'C': conserved cysteine
```

involved in a disulfide bond. '*' : position of the pattern.

Consensus pattern: G-x-[GF]-x-C-x-T-[GA]-D-C-x(1,2)-G-x(2,3)-C

[1] Edens L., Heslinga L., Klok R., Ledebore A.M., Maat J., Toonen M.Y., Visser C.,
Verrips C.T. Gene 18:1-12(1982).[2] Zhu B., Chen T.H.H., Li P.H. Plant Physiol. 108:929-
937(1995).[3] Malehorn D.E., Borgmeyer J.R., Smith C.E., Shah D.M.; Plant Physiol.
30 106:1471-1481(1994).

667. Thiolases signatures

543

Two different types of thiolase [1,2,3] are found both in eukaryotes and in prokaryotes:

acetoacetyl-CoA thiolase (EC 2.3.1.9) and 3-ketoacyl-CoA thiolase (EC 2.3.1.16). 3-ketoacyl-CoA thiolase (also called thiolase I) has a broad chain-length specificity for its substrates and is involved in degradative pathways such as fatty acid beta-oxidation. Acetoacetyl-CoA thiolase (also called thiolase II) is specific for the thiolysis of acetoacetyl-CoA and involved in biosynthetic pathways such as poly beta-hydroxybutyrate synthesis or steroid biogenesis. In eukaryotes, there are two forms of 3-ketoacyl-CoA thiolase: one located in the mitochondrion and the other in peroxisomes. There are two conserved cysteine residues important for thiolase activity. The first located in the N-terminal section of the enzymes is involved in the formation of an acyl-enzyme intermediate; the second located at the C-terminal extremity is the active site base involved in deprotonation in the condensation reaction. Mammalian nonspecific lipid-transfer protein (nsL-TP) (also known as sterol carrier protein 2) is a protein which seems to exist in two different forms: a 14 Kd protein (SCP-2) and a larger 58 Kd protein (SCP-x). The former is found in the cytoplasm or the mitochondria and is involved in lipid transport; the latter is found in peroxisomes. The C-terminal part of SCP-x is identical to SCP-2 while the N-terminal portion is evolutionary related to thiolases[4]. Three signature patterns have been developed for this family of proteins, two of which are based on the regions around the biologically important cysteines. The third is based on a highly conserved region in the C-terminal part of these proteins.

Consensus pattern: [LIVM]-[NST]-x(2)-C-[SAGLI]-[ST]-[SAG]-[LIVMFYNS]-x- [STAG]-[LIVM]-x(6)-[LIVM] [C is involved in formation of acyl-enzyme intermediate]

Consensus pattern: N-x(2)-G-G-x-[LIVM]-[SA]-x-G-H-P-x-[GA]-x-[ST]-G

Consensus pattern: [AG]-[LIVMA]-[STAGCLIVM]-[STAG]-[LIVMA]-C-x-[AG]-x-[AG]-x- [AG]-x-[SAG] [C is the active site residue]

[1] Peoples O.P., Sinskey A.J. J. Biol. Chem. 264:15293-15297(1989).[2] Yang S.-Y., Yang X.-Y.H., Healy-Louie G., Schulz H., Elzinga M. J. Biol. Chem. 265:10424-10429(1990).[3] Igual J.C., Gonzalez-Bosch C., Dopazo J., Perez-Ortin J.E. J. Mol. Evol. 35:147-155(1992).[4] Baker M.E., Billheimer J.T., Strauss J.F. III DNA Cell Biol. 10:695-698(1991).

668. Thioredoxin family active site

Thioredoxins [1 to 4] are small proteins of approximately one hundred amino-acid residues which participate in various redox reactions via the reversible oxidation of an active center

disulfide bond. They exist in either a reduced form or an oxidized form where the two cysteine residues are linked in an intramolecular disulfide bond. Thioredoxin is present in prokaryotes and eukaryotes and the sequence around the redox-active disulfide bond is wellconserved. Bacteriophage T4 also encodes for a thioredoxin but its primary structure is not homologous to bacterial, plant and vertebrate thioredoxins. A number of eukaryotic proteins contain domains evolutionary related to thioredoxin, all of them seem to be protein disulphide isomerases (PDI). PDI(EC 5.3.4.1) [5,6,7] is an endoplasmic reticulum enzyme that catalyzes the rearrangement of disulfide bonds in various proteins. The various forms of PDI which are currently known are: - PDI major isozyme; a multifunctional protein that also function as the beta subunit of prolyl 4-hydroxylase (EC 1.14.11.2), as a component of oligosaccharyl transferase (EC 2.4.1.119), as thyroxine deiodinase (EC 3.8. 1.4), as glutathione-insulin transhydrogenase (EC 1.8.4.2) and as a thyroid hormone-binding protein ! - ERp60 (ER-60; 58 Kd microsomal protein). ERp60 was originally thought to be a phosphoinositide-specific phospholipase C isozyme and later to be a protease. - ERp72. - P5. All PDI contains two or three (ERp72) copies of the thioredoxin domain. Bacterial proteins that act as thiol:disulfide interchange proteins that allows disulfide bond formation in some periplasmic proteins also contain a thioredoxin domain. These proteins are: - Escherichia coli dsbA (or prfA) and its orthologs in Vibrio cholerae (tcpG) and Haemophilus influenzae (por). - Escherichia coli dsbC (or xpRA) and its orthologs in Erwinia chrysanthemi and Haemophilus influenzae. - Escherichia coli dsbD (or dipZ) and its Haemophilus influenzae ortholog. - Escherichia coli dsbE (or ccmG) and orthologs in Haemophilus influenzae, Rhodobacter capsulatus (helX), Rhizobiaceae (cycY and tlpA). Consensus pattern: [LIVMF]-[LIVMSTA]-x-[LIVMFYC]-[FYWSTHE]-x(2)-[FYWGNTN]-C-[GATPLVE]-[PHYWSTA]-C-x(6)-[LIVMFYWT] [The two C's form the redox-active bond]

[1] Holmgren A. Annu. Rev. Biochem. 54:237-271(1985).[2] Gleason F.K., Holmgren A. FEMS Microbiol. Rev. 54:271-297(1988).[3] Holmgren A. J. Biol. Chem. 264:13963-13966(1989).[4] Eklund H., Gleason F.K., Holmgren A. Proteins 11:13-28(1991).[5] Freedman R.B., Hawkins H.C., Murrant S.J., Reid L. Biochem. Soc. Trans. 16:96-99(1988).[6] Kivirikko K.I., Myllyla R., Pihlajaniemi T. FASEB J. 3:1609-1617(1989).[7] Freedman R.B., Hirst T.R., Tuite M.F. Trends Biochem. Sci. 19:331-336(1994).

669. (Transcript fac2) Transcription factor TFIIB repeat signature

In eukaryotes the initiation of transcription of protein encoding genes by polymerase II is modulated by general and specific transcription factors. The general transcription factors operate through common promoters elements (such as the TATA box). At least seven different proteins associates to form the general transcription factors: TFIIA, -IIB, -IID, -IIE, -IIF, -IIG, and -IIH[1]. Transcription factor IIB (TFIIB) plays a central role in the transcription of class II genes, it associates with a complex of TFIID-IIA bound to DNA (DA complex) to form a ternary complex TFIID-IIA-IBB (DAB complex) which is then recognized by RNA polymerase II [2,3]. TFIIB is a protein of about 315 to 340 amino acid residues which contains, in its C-terminal part an imperfect repeat of a domain of about 75 residues. This repeat could contribute an element of symmetry to the folded protein. The following proteins have been shown to be evolutionary related to TFIIB: - An archaebacterial TFIIB homolog. In *Pyrococcus woesei* a previously undetected open reading frame has been shown [4] to be highly related to TFIIB. - Fungal transcription factor IIB 70 Kd subunit (gene PCF4/TDS4/BRF1) [5]. This protein is a general activator of RNA polymerase III transcription and plays a role analogous to that of TFIIB in pol III transcription. The central section of the repeated domain, which is the most conserved part of that domain has been selected as a signature pattern.

Consensus pattern: G-[KR]-x(3)-[STAGN]-x-[LIVMYA]-[GSTA](2)-[CSAV]-[LIVM]-[LIVMFY]-[LIVMA]-[GSA]-[STAC]

[1] Weinmann R. Gene Expr. 2:81-91(1992).[2] Hawley D. Trends Biochem. Sci. 16:317-318(1991).[3] Ha I., Lane W.S., Reinberg D. Nature 352:689-695(1991).[4] Ouzounis C., Sander C. Cell 71:189-190(1992).[5] Khoo B., Brophy B., Jackson S.P. Genes Dev. 8:2879-2890(1994).

670. (transcript fact) MADS-box domain signature and profile

A number of transcription factors contain a conserved domain of 56 amino-acid residues, sometimes known as the MADS-box domain [E1]. They are listed below: - Serum response factor (SRF) [1], a mammalian transcription factor that binds to the Serum Response Element (SRE). This is a short sequence of dyad symmetry located 300 bp to the 5' end of the transcription initiation site of genes such as c-fos. - Mammalian myocyte-specific enhancer factors 2A to 2D (MEF2A to MEF2D). These proteins are transcription factor which binds

specifically to the MEF2 element present in the regulatory regions of many muscle-specific genes. - Drosophila myocyte-specific enhancer factor 2 (MEF2). - Yeast GRM/PRTF protein (gene MCM1) [2], a transcriptional regulator of mating-type-specific genes. - Yeast arginine metabolism regulation protein I (gene ARGR1 or ARG80). - Yeast transcription factor RLM1. - Yeast transcription factor SMP1. - Arabidopsis thaliana agamous protein (AG) [3], a probable transcription factor involved in regulating genes that determines stamen and carpel development in wild-type flowers. Mutations in the AG gene result in the replacement of the stamens by petals and the carpels by a new flower. - Arabidopsis thaliana homeotic proteins Apetala1 (AP1), Apetala3 (AP3) and Pistillata (PI) which act locally to specify the identity of the floral meristem and to determine sepal and petal development [4]. - Antirrhinum majus and tobacco homeotic protein deficiens (DEFA) and globosa (GLO) [5]. Both proteins are transcription factors involved in the genetic control of flower development. Mutations in DEFA or GLO cause the transformation of petals into sepals and of stamens into carpels. - Arabidopsis thaliana putative transcription factors AGL1 to AGL6 [6]. - Antirrhinum majus morphogenetic protein DEF H33 (squamosa). In SRF, the conserved domain has been shown [1] to be involved in DNA-binding and dimerization. A pattern that spans the complete length of the domain has been derived. The profile also spans the length of the MADS-box. Consensus pattern: R-x-[RK]-x(5)-I-x-[DNGSK]-x(3)-[KR]-x(2)-T-[FY]-x-[RK](3)-x(2)-[LIVM]-x-K(2)-A-x-E-[LIVM]-[STA]-x-L-x(4)-[LIVM]-x-[LIVM](3)-x(6)-[LIVMF]-x(2)-[FY]

[1] Norman C., Runswick M., Pollock R., Treisman R. Cell 55:989-1003(1988).[2] Passmore S., Maine G.T., Elble R., Christ C., Tye B.-K. J. Mol. Biol. 204:593-606(1988).[3] Yanofsky M., Ma H., Bowman J., Drews G., Feldmann K.A., Meyerowitz E.M. Nature 346:35-39(1990).[4] Goto K., Meyerowitz E.M. Genes Dev. 8:1548-1560(1994).[5] Troebner W., Ramirez L., Motte P., Hue I., Huijser P., Loennig W.-E., Saedler H., Sommer H., Schwartz-Sommer Z. EMBO J. 11:4693-4704(1992).[6] Ma H., Yanofsky M.F., Meyerowitz E.M. Genes Dev. 5:484-495(1991).[E1]

671. Transketolase signatures

Transketolase (EC 2.2.1.1) (TK) catalyzes the reversible transfer of a two-carbon ketol unit from xylulose 5-phosphate to an aldose receptor, such as ribose 5-phosphate, to form sedoheptulose 7-phosphate and glyceraldehyde 3-phosphate. This enzyme, together with

transaldolase, provides a link between the glycolytic and pentose-phosphate pathways. TK requires thiamin pyrophosphate as a cofactor. In most sources where TK has been purified, it is a homodimer of approximately 70 Kd subunits. TK sequences from a variety of eukaryotic and prokaryotic sources [1,2] show that the enzyme has been evolutionarily conserved. In the peroxisomes of methylotrophic yeast *Hansenula polymorpha*, there is a highly related enzyme, dihydroxy-acetone synthase (DHAS) (EC 2.2.1.3) (also known as formaldehyde transketolase), which exhibits a very unusual specificity by including formaldehyde amongst its substrates. 1-deoxyxylulose-5-phosphate synthase (DXP synthase) [3] is an enzyme so far found in bacteria (gene *dxs*) and plants (gene *CLA1*) which catalyzes the thiamin pyrophosphate-dependent acyloin condensation reaction between carbon atoms 2 and 3 of pyruvate and glyceraldehyde 3-phosphate to yield 1-deoxy-D- xylulose-5-phosphate (dyp), a precursor in the biosynthetic pathway to isoprenoids, thiamin (vitamin B1), and pyridoxol (vitamin B6). DXP synthase is evolutionary related to TK. Two regions of TK have been selected as signature patterns. The first, located in the N-terminal section, contains a histidine residue which appears to function in proton transfer during catalysis [4]. The second, located in the central section, contains conserved acidic residues that are part of the active cleft and may participate in substrate-binding [4].

Consensus pattern: R-x(3)-[LIVMTA]-[DENQSTHKF]-x(5,6)-[GSN]-G-H-[PLIVMF]-[GSTA]-x(2)-[LIMC]-[GS

Consensus pattern: G-[DEQGSA]-[DN]-G-[PAEQ]-[ST]-[HQ]-x-[PAGM]-[LIVMYAC]-[DEFYW]-x(2)-[STAP]-x(2)-[RGA]

[1] Abedinia M., Layfield R., Jones S.M., Nixon P.F., Mattick J.S. *Biochem. Biophys. Res. Commun.* 183:1159-1166(1992).[2] Fletcher T.S., Kwee I.L., Nakada T., Largman C., Martin B.M. *Biochemistry* 31:1892-1896(1992).[3] Sprenger G.A., Schorken U., Wiegert T., Grolle S., De Graaf A.A., Taylor S.V., Begley T.P., Bringer-Meyer S., Sahm H. *Proc. Natl. Acad. Sci. U.S.A.* 94:12857-12862(1997).[4] Lindqvist Y., Schneider G., Ermler U., Sundstroem M. *EMBO J.* 11:2373-2379(1992).

672. Transmembrane 4 family signature

Recently a number of eukaryotic cell surface antigens have been found to be evolutionary related [1,2,3]. The proteins known to belong to this family are listed below: - Mammalian antigen CD9 (MIC3); A protein involved in platelet activation and aggregation. - Mammalian

leukocyte antigen CD37, expressed on B lymphocytes. - Mammalian leukocyte antigen CD53 (OX-44), which may be involved in growth regulation in hematopoietic cells. - Mammalian lysosomal membrane protein CD63 (melanoma-associated antigen ME491; antigen AD1). - Mammalian antigen CD81 (cell surface protein TAPA-1), which may play an important role in the regulation of lymphoma cell growth. - Mammalian antigen CD82 (protein R2; antigen C33; Kangai 1 (KAI1)), which associates with CD4 or CD8 and delivers costimulatory signals for the TCR/CD3 pathway. - Mammalian antigen CD151 (SFA-1; platelet-endothelial tetraspan antigen 3 (PETA-3)). - Mammalian cell surface glycoprotein A15 (TALLA-1; MXS1). - Mammalian novel antigen 2 (NAG-2). - Human tumor-associated antigen CO-029. - *Schistosoma mansoni* and *japonicum* 23 Kd surface antigen (SM23 / SJ23). These proteins share the following characteristics: they all seem to be type III membrane proteins (type III proteins are integral membrane proteins that contain a N-terminal membrane-anchoring domain which is not cleaved during biosynthesis and which functions both as a translocation signal and as a membrane anchor); they also contain three additional transmembrane regions, at least seven conserved cysteines residues, and are of approximately the same size (218 to 284 residues). These proteins are collectively known as the 'transmembrane 4 super family' (TM4) because they span the plasma membrane four times. A schematic diagram of the domain structure of these proteins is shown below.

+--+-----+-----+----+-----+-----+-----+-----
 -----+-----+-----+ || TMa | Extra | TM2| Cyt | TM3 | Extracellular | TM4 | Cyt| +--+-----
 +-----+----C-----C-----+-----CC-----C-----C---+-----C---+ *****Cyt : cytoplasmic domain. TMa : transmembrane anchor. TM2 to TM4: transmembrane regions 2 to 4. 'C' : conserved cysteine. '*' : position of the pattern.

A conserved region that includes two cysteines and seems to be located in a short cytoplasmic loop between two transmembrane domains has been selected as a signature for these proteins.

Consensus pattern: G-x(3)-[LIVMF]-x(2)-[GSA]-[LIVMF](2)-G-C-x-[GA]-[STA]-x(2)-[EG]-x(2)-[CWN]-[LIVM](2)

[1] Levy S., Nguyen V.Q., Andria M.L., Takahashi S. J. Biol. Chem. 266:14597-

14602(1991).[2] Tomlinson M.G., Williams A.F., Wright M.D. Eur. J. Immunol. 23:136-

40(1993).[3] Barclay A.N., Birkeland M.L., Brown M.H., Beyers A.D., Davis S.J., Somoza C., Williams A.F. The leucocyte antigen factbooks. Academic Press, London / San Diego, (1993).

673. Tryptophan synthase alpha chain signature

Tryptophan synthase catalyzes the last step in the biosynthesis of tryptophan: the conversion of indoleglycerol phosphate and serine, totryptophan and glyceraldehyde 3-phosphate [1,2]. It has two functional domains: one for the aldol cleavage of indoleglycerol phosphate to indole and glyceraldehyde 3-phosphate and the other for the synthesis of tryptophan from indole and serine. In bacteria and plants [3], each domain is found on a separate subunit (alpha and beta chains), while in fungi the two domains are fused together on a single multifunctional protein. A conserved region that contains three conserved acidic residues has been selected as a signature pattern for the alpha chain. The first and the third acidic residues are believed to serve as proton donors/acceptors in the enzyme's catalytic mechanism.

Consensus pattern: [LIVM]-E-[LIVM]-G-x(2)-[FYC]-[ST]-[DE]-[PA]-[LIVMY]-[AGLI]-[DE]-G

[1] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).[2] Hyde C.C., Miles E.W. Bio/Technology 8:27-32(1990).[3] Berlyn M.B., Last R.L., Fink G.R. Proc. Natl. Acad. Sci. U.S.A. 86:4604-4608(1989).

674. Tryptophan synthase beta chain pyridoxal-phosphate attachment site

Tryptophan synthase catalyzes the last step in the biosynthesis of tryptophan: the conversion of indoleglycerol phosphate and serine, totryptophan and glyceraldehyde 3-phosphate [1,2]. It has two functional domains: one for the aldol cleavage of indoleglycerol phosphate to indole and glyceraldehyde 3-phosphate and the other for the synthesis of tryptophan from indole and serine. In bacteria and plants [3], each domain is found on a separate subunit (alpha and beta chains), while in fungi the two domains are fused together on a single multifunctional protein. The beta chain of the enzyme requires pyridoxal-phosphate as a cofactor. The pyridoxal-phosphate group is attached to a lysine residue. The region around this lysine residue also contains two histidine residues which are part of the pyridoxal-phosphate binding site. The signature pattern for the tryptophan synthase beta chain is derived from that conserved region.

-Consensus pattern: [LIVM]-x-H-x-G-[STA]-H-K-x-N [K is the pyridoxal-P attachment site]

[1] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).[2] Hyde C.C., Miles E.W. Bio/Technology 8:27-32(1990).[3] Berlyn M.B., Last R.L., Fink G.R. Proc. Natl. Acad. Sci. U.S.A. 86:4604-4608(1989).

675. Serine proteases, trypsin family, active sites

The catalytic activity of the serine proteases from the trypsin family is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is hydrogen-bonded to a serine. The sequences in the vicinity of the active site serine and histidine residues are well conserved in this family of proteases [1]. A partial list of proteases known to belong to the trypsin family is shown below. - Acrosin. - Blood coagulation factors VII, IX, X, XI and XII, thrombin, plasminogen, and protein C. - Cathepsin G. -

Chymotrypsins. - Complement components C1r, C1s, C2, and complement factors B, D and I. - Complement-activating component of RA-reactive factor. - Cytotoxic cell proteases (granzymes A to H). - Duodenase I. - Elastases 1, 2, 3A, 3B (protease E), leukocyte (medullasin). - Enterokinase (EC 3.4.21.9) (enteropeptidase). - Hepatocyte growth factor activator. - Hepsin. - Glandular (tissue) kallikreins (including EGF-binding protein types A, B, and C, NGF-gamma chain, gamma-renin, prostate specific antigen (PSA) and tonin). - Plasma kallikrein. - Mast cell proteases (MCP) 1 (chymase) to 8. - Myeloblastin (proteinase 3) (Wegener's autoantigen). - Plasminogen activators (urokinase-type, and tissue-type). - Trypsins I, II, III, and IV. - Trypsases. - Snake venom proteases such as ancrod, batroxobin, cerastobin, flavoxobin, and protein C activator. - Collagenase from common cattle grub and collagenolytic protease from Atlantic sand fiddler crab. - Apolipoprotein(a). - Blood fluke cercarial protease. - Drosophila trypsin like proteases: alpha, easter, snake-locus. - Drosophila protease stubble (gene sb). - Major mite fecal allergen Der p III. All the above proteins belong to family S1 in the classification of peptidases[2,E1] and originate from eukaryotic species. It should be noted that bacterial proteases that belong to family S2A are similar

enough in the regions of the active site residues that they can be picked up by the same patterns. These proteases are listed below. - Achromobacter lyticus protease I. - Lysobacter alpha-lytic protease. - Streptogrisin A and B (Streptomyces proteases A and B). -

Streptomyces griseus glutamyl endopeptidase II. - Streptomyces fradiae proteases 1 and 2. Consensus pattern: [LIVM]-[ST]-A-[STAG]-H-C [H is the active site residue]

Consensus pattern: [DNSTAGC]-[GSTAPIMVQH]-x(2)-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]-[LIVMFYSTANQH] [S is the active site residue]

[1] Brenner S. Nature 334:528-530(1988).[2] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

676. (tsp) Thrombospondin type 1 domain

5 [1] Bork P; FEBS lett 1993;327:125-130.

677. Tubulin subunits alpha, beta, and gamma signature

10 Tubulins [1,2], the major constituent of microtubules are dimeric proteins which consist of two closely related subunits (alpha and beta). Tubulin binds two molecules of GTP at two different sites (N and E). At the E (Exchangeable) site, GTP is hydrolyzed during incorporation into the microtubule. Near the E site is an invariant region rich in glycines which is found in both chains and which is now [3] said to control the access of the nucleotide to its binding site. A signature pattern was developed from this region. With the exception of the simple eukaryotes, most species express a variety of closely related alpha and beta isotypes. In most species there is a third member of the tubulin family: gamma tubulin. Gamma tubulin is found at microtubule organizing centers (MTOC) such as the spindle poles or the centrosome, suggesting that it is involved in the minus-end nucleation of microtubule assembly [4].

15
20 Consensus pattern: [SAG]-G-G-T-G-[SA]-G

[1] Cleveland D.W., Sullivan K.F. Annu. Rev. Biochem. 54:331-365(1985).[2] Joshi H.C., Cleveland D.W. Cell Motil. Cytoskeleton 16:159-163(1990).[3] Hesse J., Thierauf M., Ponstingl H. J. Biol. Chem. 262:15472-15475(1987).[4] Joshi H.C. BioEssays 15:637-643(1993).

25 Tubulin-beta mRNA autoregulation signal

30 The stability of beta-tubulin mRNAs are autoregulated by their own translation product [1]. Unpolymerized tubulin subunits bind directly (or activate a factor(s) which binds co-translationally) to the nascent N-terminus of beta-tubulin. This binding is transduced through the adjacent ribosomes to activate an RNase that degrades the polysome-bound mRNA. The recognition element has been shown to be the first four amino acids of beta-tubulin: Met-Arg-Glu-Ile. Mutations to this sequence abolish the autoregulation effect (except for the

replacement of Glu by Asp); transposition of this sequence to an internal region of a polypeptide also suppresses the autoregulatory effect.

Consensus pattern: <M-R-[DE]-[IL]

[1] Cleveland D.W. Trends Biochem. Sci. 13:339-343(1988).

5

678. (tRNA-synt 2c) Aminoacyl-transfer RNA synthetases class-II signatures. Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]-

Consensus pattern: [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]-

25

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D. BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S. Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

30

679. UBA-domain

The UBA-domain (ubiquitin associated domain) is a novel sequence motif found in several proteins having connections to ubiquitin and the ubiquitination pathway. The structure of the UBA domain consists of a compact three helix bundle [1]. Number of members: 84

[1] Structure of a human DNA repair protein UBA domain that interacts with HIV-1 Vpr. Dieckmann T, Withers-Ward ES, Jarosinski MA, Liu CF, Chen IS, Feigon J; Nat Struct Biol 1998;5:1042-1047.

680. UBX domain

Domain present in ubiquitin-regulatory proteins. Present in FAF1 and Shp1p. Number of members: 19

[1] The UBA domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway. Hofmann K, Bucher P; Trends Biochem Sci 1996;21:172-173.

681. (UCH) Ubiquitin carboxyl-terminal hydrolases family 1 cysteine active site

Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well as that of ubiquitinated proteins. There are two distinct families of UCH. The first class consist of enzymes of about 25 Kd and is currently represented by: - Mammalian isozymes L1 and L3. - Yeast YUH1. - Drosophila Uch. One of the active site residues of class-I UCH [3] is a cysteine. A signature pattern has been derived from the region around that residue. Consensus pattern: Q-x(3)-N-[SA]-C-G-x(3)-[LIVM](2)-H-[SA]-[LIVM]-[SA] [C is the active site residue

[1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).[2] D'andrea A., Pellman D. Crit. Rev. Biochem. Mol. Biol. 33:337-352(1998).[3] Johnston S.C., Larsen C.N., Cook W.J., Wilkinson K.D., Hill C.P. EMBO J. 16:3787-3796(1997).[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

682. Ubiquitin carboxyl-terminal hydrolases family 2 signatures (UCH-1)

Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well as that of ubiquitinated proteins. There are two distinct families of UCH. The second class consist of largeproteins (800 to 2000 residues) and is currently represented by: - Yeast UBP1, UBP2, UBP3, UBP4 (or DOA4/SSV7), UBP5, UBP7, UBP9, UBP10, UBP11, UBP12, UBP13, UBP14, UBP15 and UBP16. - Human tre-2. - Human isopeptidase T. - Human isopeptidase T-3. - Mammalian Ode-1. - Mammalian Unp. - Mouse Dub-1. - Drosophila fat facets protein (gene *faf*). - Mammalian *faf* homolog. - Drosophila D-Ubp-64E. - *Caenorhabditis elegans* hypothetical protein R10E11.3. - *Caenorhabditis elegans* hypothetical protein K02C4.3. These proteins only share two regions of similarity. The first region contains a conserved cysteine which is probably implicated in the catalytic mechanism. The second region contains two conserved histidines residues, one of which is also probably implicated in the catalytic mechanism. Signature patterns for both conserved regions have been developed.

Consensus pattern: G-[LIVMFY]-x(1,3)-[AGC]-[NASM]-x-C-[FYW]-[LIVMC]-[NST]-[SACV]-x-[LIVMS]-Q [C is the putative active site residue]

Consensus pattern: Y-x-L-x-[SAG]-[LIVMFT]-x(2)-H-x-G-x(4,5)-G-H-Y [The two H's are putative active site residues]

[1] Jentsch S., Seufert W., Hauser H.-P. *Biochim. Biophys. Acta* 1089:127-139(1991).[2] D'andrea A., Pellman D. *Crit. Rev. Biochem. Mol. Biol.* 33:337-352(1998).[3] Rawlings N.D., Barrett A.J. *Meth. Enzymol.* 244:461-486(1994).

683. Ubiquitin carboxyl-terminal hydrolases family 2 signatures (UCH-2)

Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well as that of ubiquitinated proteins. There are two distinct families of UCH. The second class consist of largeproteins (800 to 2000 residues) and is currently represented by: - Yeast UBP1, UBP2, UBP3, UBP4 (or DOA4/SSV7), UBP5, UBP7, UBP9, UBP10, UBP11, UBP12,

UBP13, UBP14, UBP15 and UBP16. - Human tre-2. - Human isopeptidase T. - Human isopeptidase T-3. - Mammalian Ode-1. - Mammalian Unp. - Mouse Dub-1. - Drosophila fat facets protein (gene faf). - Mammalian faf homolog. - Drosophila D-Ubp-64E. - Caenorhabditis elegans hypothetical protein R10E11.3. - Caenorhabditis elegans hypothetical protein K02C4.3. These proteins only share two regions of similarity. The first region contains a conserved cysteine which is probably implicated in the catalytic mechanism. The second region contains two conserved histidines residues, one of which is also probably implicated in the catalytic mechanism. Signature patterns for both conserved regions have been developed.

Consensus pattern: G-[LIVMFY]-x(1,3)-[AGC]-[NASM]-x-C-[FYW]-[LIVMC]-[NST]-[SACV]-x-[LIVMS]-Q [C is the putative active site residue]

Consensus pattern: Y-x-L-x-[SAG]-[LIVMFT]-x(2)-H-x-G-x(4,5)-G-H-Y [The two H's are putative active site residues]

[1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991). [2] D'andrea A., Pellman D. Crit. Rev. Biochem. Mol. Biol. 33:337-352(1998). [3] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

684. UDP-glycosyltransferases signature

UDP glycosyltransferases (UGT) are a superfamily of enzymes that catalyzes the addition of the glycosyl group from a UTP-sugar to a small hydrophobic molecule. This family currently consist of: - Mammalian UDP-glucuronosyl transferases (UDPGT) [1,2]. A large family of membrane-bound microsomal enzymes which catalyze the transfer of glucuronic acid to a wide variety of exogenous and endogenous lipophilic substrates. These enzymes are of major importance in the detoxification and subsequent elimination of xenobiotics such as drugs and carcinogens. - A large number of putative UDPGT from Caenorhabditis elegans. - Mammalian 2-hydroxyacylsphingosine 1-beta-galactosyltransferase [3] (also known as UDP-galactose-ceramide galactosyltransferase). This enzyme catalyzes the transfer of galactose to ceramide, a key enzymatic step in the biosynthesis of galactocerebrosides, which are abundant sphingolipids of the myelin membrane of the central nervous system and peripheral nervous system. - Plants flavonol O(3)-glucosyltransferase. An enzyme [4] that catalyzes the transfer of glucose from UDP-glucose to a flavanol. This reaction is essential and one of the last steps in anthocyanin pigment biosynthesis. - Baculoviruses ecdysteroid UDP-

glucosyltransferase (EC 2.4.1.-) [5] (egt). This enzyme catalyzes the transfer of glucose from UDP-glucose to ectysteroids which are insect molting hormones. The expression of egt in the insect host interferes with the normal insect development by blocking the molting process. -

Prokaryotic zeaxanthin glucosyl transferase (gene crtX), an enzyme involved in carotenoid biosynthesis and that catalyses the glycosylation reaction which converts zeaxanthin to zeaxanthin-beta- diglucoside. - Streptomyces macrolide glycosyltransferases [6]. These enzymes specifically inactivates macrolide antibiotics via 2'-O-glycosylation using UDP-glucose. These enzymes share a conserved domain of about 50 amino acid residues located in their C-terminal section and from which a pattern has been extracted to detect them.

Consensus pattern: [FW]-x(2)-Q-x(2)-[LIVMYA]-[LIMV]-x(4,6)-[LVGAC]-[LVFYA]-[LIVMF]-[STAGCM]-[HNQ]-[STAGC]-G-x(2)-[STAG]-x(3)-[STAGL]-[LIVMFA]-x(4)-[PQR]-[LIVMT]-x(3)-[PA]-x(3)-[DES]-[QEHN]

[1] Dutton G.J. (In) Glucoronidation of drugs and other compounds, Dutton G.J., Ed., pp 1-78, CRC Press, Boca Raton, (1980).[2] Burchell B., Nebert D.W., Nelson D.R., Bock K.W., Iyanagi T., Jansen P.L., Lancet D., Mulder G.J., Chowdhury J.R., Siest G., Tephly T.R., Mackenzie P.I. DNA Cell Biol. 10:487-494(1991).[3] Schulte S., Stoffel W. Proc. Natl. Acad. Sci. U.S.A. 90:10265-10269(1993).[4] Furtek D., Schiefelbein J.W., Johnston F., Nelson O.E. Jr. Plant Mol. Biol. 11:473-481(1988).[5] O'Reilly D.R., Miller L.K. Science 245:1110-1112(1989).[6] Hernandez C., Olano C., Mendez C., Salas J.A. Gene 134:139-140(1993).

685. UDP-glucose/GDP-mannose dehydrogenase family

The UDP-glucose/GDP-mannose dehydrogenases are a small group of enzymes which possesses the ability to catalyze the NAD-dependent 2-fold oxidation of an alcohol to an acid without the release of an aldehyde intermediate [2]. Number of members: 55

[1] Purification and characterization of guanosine diphospho-D-mannose dehydrogenase. A key enzyme in the biosynthesis of alginate by *Pseudomonas aeruginosa*. Roychoudhury S, May TB, Gill JF, Singh SK, Feingold DS, Chakrabarty AM; J Biol Chem 1989;264:9380-9385. [2] Properties and kinetic analysis of UDP-glucose dehydrogenase from group A streptococci. Irreversible inhibition by UDP-chloroacetol. Campbell RE, Sala RF, van de Rijn I, Tanner ME; J Biol Chem 1997;272:3416-3422.

686. Uracil-DNA glycosylase signature

Uracil-DNA glycosylase (EC 3.2.2.-) (UNG) [1] is a DNA repair enzyme that excises uracil residues from DNA by cleaving the N-glycosylic bond. Uracil in DNA can arise as a result of misincorporation of dUMP residues by DNA polymerase or deamination of cytosine. The sequence of uracil-DNA glycosylase is extremely well conserved [2] in bacteria and eukaryotes as well as in herpes viruses. More distantly related uracil-DNA glycosylases are also found in poxviruses [3]. In eukaryotic cells, UNG activity is found in both the nucleus and the mitochondria. Human UNG1 protein is transported to both the mitochondria and the nucleus [4]. The N-terminal 77 amino acids of UNG1 seem to be required for mitochondrial localization [4], but the presence of a mitochondrial transit peptide has not been directly demonstrated. As a signature for this type of enzyme, the most N-terminal conserved region has been selected. This region contains an aspartic acid residue which has been proposed, based on X-ray structures [5,6] to act as a general base in the catalytic mechanism. Consensus pattern: [KR]-[LIV]-[LIVC]-[LIVM]-x-G-[QI]-D-P-Y [D is the active site residue]-

[1] Sancar A., Sancar G.B. *Annu. Rev. Biochem.* 57:29-67(1988).[2] Olsen L.C., Aasland R., Wittwer C.U., Krokan H.E., Helland D.E. *EMBO J.* 8:3121-3125 (1989).[3] Upton C., Stuart D.T., McFadden G. *Proc. Natl. Acad. Sci. U.S.A.* 90:4518-4522(1993).[4] Slupphaug G., Markussen F.-H., Olsen L.C., Aasland R., Aarsaether N., Bakke O., Krokan H.E., Helland D.E. *Nucleic Acids Res.* 21:2579-2584(1993).[5] Savva R., McAuley-Hecht K., Brown T., Pearl L. *Nature* 373:487-493(1995).[6] Mol C.D., Arvai A.S., Slupphaug G., Kavli B., Alseth I., Krokan H.E., Tainer J.A. *Cell* 80:869-878(1995).[7] Muller S.J., Caradonna S. *Biochim. Biophys. Acta* 1088:197-207(1991).[8] Meyer-Siegler K., Mauro D.J., Seal G., Wurzer J., Deriel J.K., Sirover M.A. *Proc. Natl. Acad. Sci. U.S.A.* 88:8460-8464(1991).[9] Muller S.J., Caradonna S. *J. Biol. Chem.* 268:1310-1319(1993).[10] Barnes D.E., Lindahl T., Sedgwick B. *Curr. Opin. Cell Biol.* 5:424-433(1993).

687. Uncharacterized protein family UPF0001 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast chromosome II hypothetical protein YBL036c. - *Caenorhabditis elegans* hypothetical protein F09E5.8. - *Bacillus subtilis* hypothetical protein ylmE. - *Escherichia coli* hypothetical

558

protein yggS and HI0090, the corresponding Haemophilus influenzae protein. - Helicobacter pylori hypothetical protein HP0395. - Mycobacterium tuberculosis hypothetical protein MtCY270.20. - Synechocystis strain PCC 6803 hypothetical protein slr0556. - A Pseudomonas aeruginosa hypothetical protein in pilT 5'region. - A Vibrio alginolyticus hypothetical protein in pilT 5'region. These are proteins of from 25 to 30 Kd which contain a number of conserved regions. The best conserved region which is located in the first third of these proteins has been selected as a signature pattern.

Consensus pattern: [FW]-H-[FM]-[IV]-G-x-[LIV]-Q-x-[NKR]-K-x(3)-[LIV]

[1] Bairoch A., Rudd K.E. Unpublished observations (1996).

10

688. Uncharacterized protein family UPF0003 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli protein aefA. - Escherichia coli hypothetical protein yggB. - Escherichia coli hypothetical protein yjeP and HI0195.1, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein ynaI. - Bacillus subtilis hypothetical protein yhdY. - Helicobacter pylori hypothetical protein HP0415. - Synechocystis strain PCC 6803 hypothetical protein slr0639. - Archaeoglobus fulgidus hypothetical protein AF1546. - Methanococcus jannaschii hypothetical protein MJ0170. - Methanococcus jannaschii hypothetical protein MJ1143. The size of these proteins range from 30 to 120 Kd. They all contain a number of transmembrane regions. The best conserved region which is located in and just after the last potential transmembrane region has been selected as a signature pattern,.

Consensus pattern: G-[STIF]-V-x(2)-[LIVM]-x(6)-[LIVMF]-x(3)-[DQ]-x(3)-[LIV]- x-[LIV]-

P-N-x(2)-[LIVMF]-[LIVFSTA]-x(5)-N

[1] Bairoch A. Unpublished observations (1997).

689. Uncharacterized protein family UPF0004 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein yliG. - Escherichia coli hypothetical protein yleA and HI0019, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yqeV. - Helicobacter pylori hypothetical protein HP0269. - Helicobacter pylori

hypothetical protein HP0285. - Mycoplasma iowae hypothetical protein in 16S RNA 5'region. - Mycobacterium leprae hypothetical protein B2235_C2_195. - Pseudomonas aeruginosa hypothetical protein in hemL 3'region. - Synechocystis strain PCC 6803 hypothetical protein slr0082. - Synechocystis strain PCC 6803 hypothetical protein slI0996. - Methanococcus jannaschii hypothetical protein MJ0865. - Methanococcus jannaschii hypothetical protein MJ0867. - Caenorhabditis elegans hypothetical protein F25B5.5. The size of these proteins range from 47 to 61 Kd. They contain six conserved cysteines, three of which are clustered in a region that can be used as a signature pattern.

Consensus pattern: [LIVM]-x-[LIVMT]-x(2)-G-C-x(3)-C-[STAN]-[FY]-C-x-[LIVM]-x(4)-

G

[1] Bairoch A. Unpublished observations (1997).

690. Uncharacterized protein family UPF0005 signature

The following proteins seem to be evolutionarily related [1]: - Mammalian protein TEGT (Testis Enhanced Gene Transcript). - Escherichia coli hypothetical protein yccA and HI0044, the corresponding Haemophilus influenzae protein. - A probable Pseudomonas aeruginosa ortholog of yccA. These are proteins of about 25 Kd which seem to contain seven transmembrane domains. A signature pattern that corresponds to a region that starts with the beginning of the third transmembrane domain and ends in the middle of the fourth one has been developed.

Consensus pattern: G-[LIVM](2)-[SA]-x(5,8)-G-x(2)-[LIVM]-G-P-x-L-x(4)-[SAG]-x(4,6)-[LIVM](2)-x(2)-A-x(3)-T-A-[LIVM](2)-F

[1] Walter L., Marynen P., Szpirer J., Levan G., Guenther E. Genomics 28:301-304(1995).

691. Uncharacterized protein family UPF0006 signatures

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast chromosome II hypothetical protein YBL055c. - Escherichia coli hypothetical protein ycfH and HI0454, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yigW. - Escherichia coli hypothetical protein yjjV and HI0081, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yabD. - Haemophilus influenzae hypothetical protein HI1664. - Mycoplasma genitalium

hypothetical protein MG009. These are proteins of from 24 to 47 Kd which contain a number of conserved regions. They can be picked up in the database by the following patterns.

Consensus pattern: [LIVMFY](2)-D-[STA]-H-x-H-[LIVMF]-[DN

Consensus pattern: P-[LIVM]-x-[LIVM]-H-x-R-x-[TA]-x-[DE

5 Consensus pattern: [LVSA]-[LIVA]-x(2)-[LIVM]-[PS]-x(3)-L-[LIVM]-[LIVMS]-E-T- D-x-P

[1] Bairoch A., Rudd K.E. Unpublished observations (1995).

10 692. Uncharacterized protein family UPF0007 signature

The following proteins seems to be evolutionary related [1]: - Escherichia coli hypothetical protein ygbP and HI0672, the corresponding Haemophilus influenzae protein. - Bacillus

subtilis hypothetical protein yacM. - Mycobacterium tuberculosis hypothetical protein

MtCY06G11.29c. - Synechocystis strain PCC 6803 hypothetical protein slr0951. - A

15 Rhodobacter capsulatus hypothetical protein in nifR3 5'region. Except for the Rhodobacter protein which contains a C-terminal extension, all these proteins have from 225 to 236 amino acids. They are hydrophilic proteins that can be picked up in the database by the following pattern.

Consensus pattern: V-L-[IV]-H-D-[GA]-A-R

20 [1] Bairoch A. Unpublished observations (1997).

693. Uncharacterized protein family UPF0015 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

25 Yeast chromosome II hypothetical protein YBR002c. - Yeast chromosome XIII hypothetical protein YMR101c. - Escherichia coli hypothetical protein yaeU and HI0920, the

corresponding Haemophilus influenzae protein. - Helicobacter pylori hypothetical protein

HP1221. - Mycobacterium leprae hypothetical protein B1937_F2_65. - A Corynebacterium

glutamicum hypothetical protein in aroF 3'region. - A Streptomyces fradiae hypothetical

30 protein in transposon Tn4556. - Synechocystis strain PCC 6803 hypothetical protein slI0505.

- Methanococcus jannaschii hypothetical protein MJ1372. These are proteins of about 26 to 40 Kd whose central region is well conserved. They can be picked up in the database by the following pattern.

561

Consensus pattern: [DE]-[LIVMF](3)-R-T-[SG]-G-x(2)-R-x-S-x-[FY]-[LIVM](2)-W-Q-
[1] Wolfe K.H., Lohan A.J.E. Yeast 10:S41-S46(1994).

5 694. Uncharacterized protein family UPF0016 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -
Yeast hypothetical protein YBR187w. - Fission yeast hypothetical protein SpAC17G8.08c. -
Mouse protein pFT27. - Synechocystis strain PCC 6803 hypothetical protein slI0615. These
are hydrophobic proteins of 200 to 320 amino acids that seem to contain six or seven
transmembrane domains. A conserved region which seems, in the eukaryotic proteins of this
family, to directly follow the second transmembrane domain has been selected as a signature
pattern.

Consensus pattern: E-[LIVM]-G-D-K-T-F-[LIVMF](2)-A-
[1] Bairoch A. Unpublished observations (1996).

695. Uncharacterized protein family UPF0021 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -
Yeast chromosome VII hypothetical protein YGL211w. - Dictyostelium discoideum protein
veg136. - Methanococcus jannaschii hypothetical proteins MJ1157 and MJ1478. These are
proteins of from 300 to 360 residues. They can be picked up in the database by the following
pattern which is located in their N-terminal section.

Consensus pattern: C-K-x(2)-F-x(4)-E-x(22,23)-S-G-G-K-D
[1] Bairoch A. Unpublished observations (1997).

696. Uncharacterized protein family UPF0023 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -
Mouse protein 22A3. - Yeast chromosome XII hypothetical protein YLR022c. -
Caenorhabditis elegans hypothetical protein W06E11.4. - Methanococcus jannaschii
hypothetical protein MJ0592. These are hydrophilic proteins of about 30 Kd. They can be
picked up in the database by the following pattern.

Consensus pattern: D-x-D-E-[LIV]-L-x(4)-V-F-x(3)-S-K-G-

[1] Bairoch A. Unpublished observations (1997).

697. Uncharacterized protein family UPF0024 signature. The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein ygbO and HI0701, the corresponding Haemophilus influenzae protein. - Helicobacter pylori hypothetical protein HP0926. - Yeast chromosome XV hypothetical protein YOR243c. - Caenorhabditis elegans hypothetical protein B0024.11. - Methanococcus jannaschii hypothetical proteins MJ0588 and MJ1364. These are hydrophilic proteins of from 39 to 77 Kd. They can be picked up in the database by the following pattern.

Consensus pattern: G-x-K-D-[KR]-x-A-[LV]-T-x-Q-x-[LIVF]-[SGC]-

[1] Bairoch A. Unpublished observations (1997).

698. Uncharacterized protein family UPF0025 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein yfcE. - Bacillus subtilis hypothetical protein ysnB. - Mycoplasma genitalium and pneumoniae hypothetical protein MG207. - Methanococcus jannaschii hypothetical proteins MJ0623 and MJ0936. These are hydrophilic proteins of about 20 Kd. They can be picked up in the database by the following pattern.

Consensus pattern: D-V-[LIV]-x(2)-G-H-[ST]-H-x(12)-[LIVMF]-N-P-G

[1] Bairoch A. Unpublished observations (1997).

699. Uncharacterized protein family UPF0029 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast chromosome III hypothetical protein YCR59c. - Yeast chromosome IV hypothetical protein YDL177C. - Escherichia coli hypothetical protein yigZ and HI0722, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yvyE. - A Thermus aquaticus hypothetical protein in pol 5' region. These proteins can be picked up in the database by the following pattern.

Consensus pattern: G-x(2)-[LIVM](2)-x(2)-[LIVM]-x(4)-[LIVM]-x(5)-[LIVM](2)-x- R-
[FYW](2)-G-G-x(2)-[LIVM]-G

[1] Koonin E.V., Bork P., Sander C. EMBO J. 13:493-503(1994).

5

700. Uncharacterized protein family UPF0030 signature

The following uncharacterized proteins have been shown [1] to be highly similar: - Yeast chromosome VI hypothetical protein YFL060c. - Yeast chromosome XIII hypothetical protein YMR095c. - Yeast chromosome XIV hypothetical protein YNL334c. - *Bacillus subtilis* hypothetical protein yaaE. - *Haemophilus influenzae* hypothetical protein HI1648. - *Methanococcus jannaschii* hypothetical protein MJ1661. These are hydrophilic proteins of about 19 to 25 Kd. They can be picked up in the database by the following pattern.

10

Consensus pattern: [GA]-L-I-[LIV]-P-G-G-E-S-T-[STA]

[1] Bairoch A. Unpublished observations (1997).

15

701. Uncharacterized protein family UPF0032 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - *Escherichia coli* hypothetical protein yigU and HI0188, the corresponding *Haemophilus influenzae* protein. - *Bacillus subtilis* hypothetical protein ycbT. - *Mycobacterium tuberculosis* hypothetical protein MtCY49.33c and U2126A, the corresponding *Mycobacterium leprae* protein. - *Synechocystis* strain PCC 6803 hypothetical protein sll0194. - *Odontella sinensis* and *Porphyra purpurea* chloroplast hypothetical protein ycf43. These proteins have from 245 to 317 amino acids and seem to contain at least six or seven transmembrane regions. A conserved region located in the central section of these proteins has been developed as a signature pattern.

20

Consensus pattern: Y-x(2)-F-[LIVMA](2)-x-L-x(4)-G-x(2)-F-[EQ]-[LIVMF]-P- [LIVM] -

[1] Bairoch A., Rudd K.E. Unpublished observations (1996).

30

702. Uncharacterized protein family UPF0034 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - *Escherichia coli* hypothetical protein yhdG and HI0979, the corresponding *Haemophilus*

influenzae protein. - Escherichia coli hypothetical protein yjbN and HI0634, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yohI and HI0270, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yacF. - Rhodobacter capsulatus protein nifR3 and related proteins in Azospirillum brasilense and Rhizobium leguminosarum. - Synechocystis strain PCC 6803 hypothetical protein slr0644. - Synechocystis strain PCC 6803 hypothetical protein slI0926. - Caenorhabditis elegans hypothetical protein C45G9.2. - Yeast protein SMM1. - Yeast hypothetical protein YLR401c. - Yeast hypothetical protein YLR405w. - Yeast hypothetical protein YML080w. Although it has been proposed [2] that Rhodobacter capsulatus nifR3 is a transcriptional regulatory protein, it is believed that these proteins constitute a family of enzymes whose active site could include a conserved cysteine which has been used as the central part of a signature pattern.

Consensus pattern: [LIVM]-[DNG]-[LIVM]-N-x-G-C-P-x(3)-[LIVMASQ]-x(5)-G-[SAC]
[1] Bairoch A., Rudd K.E. Unpublished observations (1995).[2] Foster-Hartnett D., Cullen P.J., Gabbert K.K., Kranz R.G. Mol. Microbiol. 8:903-914(1993).

703. Uncharacterized protein family UPF0038 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein yacE and HI0890, the corresponding Haemophilus influenzae protein. - Mycobacterium tuberculosis hypothetical protein MtCY01B2.23 and O410, the corresponding Mycobacterium leprae protein. - Synechocystis strain PCC 6803 hypothetical protein slr0553. - Other hypothetical proteins from Aeromonas hydrophila, Bacteroides nodosus, Neisseria gonorrhoeae, Pseudomonas putida, Thermus thermophilus and Xanthomonas campestris. - Human hypothetical protein pOV-2. - Yeast hypothetical protein YDR196C. - Caenorhabditis elegans hypothetical protein T05G5.5. These proteins all contain, in their N-terminal extremity, an ATP/GTP-binding motif 'A' (P-loop) (see <PDOC00017>). The size of these proteins range from 200 to 290 residues (with the exception of the Mycobacterial sequences which are 410 residues long). A conserved region some 50 residues away from the ATP-binding P-loop has been developed as a signature pattern.

Consensus pattern: G-x-[LI]-x-R-x(2)-L-x(4)-F-x(8)-[LIV]-x(5)-P-x-[LIV] -
[1] Rudd K.E., Bairoch A. Unpublished observations (1997).

704. Ubiquitin-conjugating enzymes active site

Ubiquitin-conjugating enzymes (UBC or E2 enzymes) [1,2,3] catalyze the covalent attachment of ubiquitin to target proteins. An activated ubiquitin moiety is transferred from an ubiquitin-activating enzyme (E1) to E2 which later ligates ubiquitin directly to substrate proteins with or without the assistance of 'N-end' recognizing proteins (E3). In most species there are many forms of UBC (at least 9 in yeast) which are implicated in diverse cellular functions. A cysteine residue is required for ubiquitin-thiolester formation. There is a single conserved cysteine in UBC's and the region around that residue is conserved in the sequence of known UBC isozymes. That region has been used as a signature pattern.

Consensus pattern: [FYWLSP]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-[LIV]-x-[LIV] [C is the active site residue]

[1] Jentsch S., Seufert W., Sommer T., Reins H.-A. Trends Biochem. Sci. 15:195-198(1990). [2] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991). [3] Hershko A. Trends Biochem. Sci. 16:265-268(1991).

705. Uroporphyrinogen decarboxylase signatures

Uroporphyrinogen decarboxylase (URO-D), the fifth enzyme of the heme biosynthetic pathway, catalyzes the sequential decarboxylation of the four acetyl side chains of uroporphyrinogen to yield coproporphyrinogen [1]. URO-D deficiency is responsible for the Human genetic diseases familial porphyria cutanea tarda (fPCT) and hepatoerythropoietic porphyria (HEP). The sequence of URO-D has been well conserved throughout evolution.

The best conserved region is located in the N-terminal section; it contains a perfectly conserved hexapeptide. There are two arginine residues in this hexapeptide which could be involved in the binding, via salt bridges, to the carboxyl groups of the propionate side chains of the substrate. This region has been used as a signature pattern. A second signature pattern is based on another well conserved region which is located in the central section of the protein.

Consensus pattern: P-x-W-x-M-R-Q-A-G-R

Consensus pattern: G-F-[STAGCV]-[STAGC]-x-P-[FYW]-T-[LV]-x(2)-Y-x(2)-[AE]-[GK]

[1] Garey J.R., Labbe-Bois R., Chelstowska A., Rytka J., Harrison L., Kushner J., Labbe P.
Eur. J. Biochem. 205:1011-1016(1992).

5 706. ubiE/COQ5 methyltransferase family signatures

The following methyltransferases have been shown [1] to share regions of similarities: -

Escherichia coli ubiE, which is involved in both ubiquinone and menaquinone biosynthesis
and which catalyzes the S-adenosylmethionine dependent methylation of 2-polyprenyl-6-
methoxy-1,4-benzoquinol into 2-polyprenyl-3- methyl-6-methoxy-1,4-benzoquinol and of

10 demethylmenaquinol into menaquinol. - Yeast COQ5, a ubiquinone biosynthesis
methlytransferase. - Bacillus subtilis spore germination protein C2 (gene: gercB or gerC2), a
probable menaquinone biosynthesis methlytransferase. - Lactococcus lactis gerC2 homolog. -
Caenorhabditis elegans hypothetical protein ZK652.9. - Leishmania donovani amastigote-
specific protein A41. These are hydrophilic proteins of about 30 Kd (except for ZK652.9
15 which is 65Kd). They can be picked up in the database by the following patterns.

Consensus pattern: Y-D-x-M-N-x(2)-[LIVM]-S-x(3)-H-x(2)-W

Consensus pattern: R-V-[LIVM]-K-[PV]-G-G-x-[LIVMF]-x(2)-[LIVM]-E-x-S

[1] Lee P.T., Hsu A.Y., Ha H.T., Clarke C.F. J. Bacteriol. 179:1748-1754(1997).

20 707. Uricase signature

Uricase (urate oxidase) [1] is the peroxisomal enzyme responsible for the degradation of
urate into allantoin. Some species, like primates and birds, have lost the gene for uricase and
are therefore unable to degradeurate. Uricase is a protein of 300 to 400 amino acids. A highly
25 conserved region located in the central part of the sequence has been used as a signature
pattern.

Consensus pattern: [LV]-x-[LV]-[LIV]-K-[STV]-[ST]-x-[SN]-x-F-x(2)-[FY]-x(4)- [FY]-
x(2)-L-x(5)-R

[1] Motojima K., Kanaya S., Goto S. J. Biol. Chem. 263:16677-16681(1988).

30 708. Universal stress protein family (Usp)

By a wide range of stress conditions members of the Usp family are predicted to be related to the MADS-box proteins transcript_fact and bind to DNA [2]. Number of members: 39

- 5 [1] Expression and role of the universal stress protein, UspA, of Escherichia coli during growth arrest. Nystrom T, Neidhardt FC; Mol Microbiol 1994; 11:537-544.
[2] Sequence analysis of eukaryotic developmental proteins: ancient and novel domains. Mushegian AR, Koonin EV; Genetics 1996; 144:817-828.

10

709. Ubiquitin domain signature and profile

Ubiquitin [1,2,3] is a protein of seventy six amino acid residues, found in all eukaryotic cells and whose sequence is extremely well conserved from protozoan to vertebrates. It plays a key role in a variety of cellular processes, such as ATP-dependent selective degradation of cellular proteins, maintenance of chromatin structure, regulation of gene expression, stress response and ribosome biogenesis. In most species, there are many genes coding for ubiquitin. However they can be classified into two classes. The first class produces polyubiquitin molecules consisting of exact head to tail repeats of ubiquitin. The number of repeats is variable (up to twelve in a Xenopus gene). In the majority of polyubiquitin precursors, there is a final amino-acid after the last repeat. The second class of genes produces precursor proteins consisting of a single copy of ubiquitin fused to a C-terminal extension protein (CEP). There are two types of CEP proteins and both seem to be ribosomal proteins. Ubiquitin is a globular protein, the last four C-terminal residues (Leu-Arg- Gly-Gly) extending from the compact structure to form a 'tail', important for its function. The latter is mediated by the covalent conjugation of ubiquitin to target proteins, by an isopeptide linkage between the C-terminal glycine and the epsilon amino group of lysine residues in the target proteins. There are a number of proteins which are evolutionary related to ubiquitin: - Ubiquitin-like proteins from baculoviruses as well as in some strains of bovine viral diarrhea viruses (BVDV). These proteins are highly similar to their eukaryotic counterparts. - Mammalian protein GDX [4]. GDX is composed of two domains, a N-terminal ubiquitin-like domain of 74 residues and a C-terminal domain of 83 residues with some similarity with the thyroglobulin hormonogenic site. - Mammalian protein FAU [5]. FAU is a fusion protein which consist of a N-terminal ubiquitin-like protein of 74 residues fused to ribosomal protein

30

S30. - Mouse protein NEDD-8 [6], a ubiquitin-like protein of 81 residues. - Human protein BAT3, a large fusion protein of 1132 residues that contains a N-terminal ubiquitin-like domain. - *Caenorhabditis elegans* protein ubl-1 [7]. Ubl-1 is a fusion protein which consist of a N-terminal ubiquitin-like protein of 70 residues fused to ribosomal protein S27A. - Yeast DNA repair protein RAD23 [8]. RAD23 contains a N-terminal domain that seems to be distantly, yet significantly, related to ubiquitin. - Mammalian RAD23-related proteins RAD23A and RAD23B. - Mammalian BCL-2 binding athanogene-1 (BAG-1). BAG-1 is a protein of 274 residues that contains a central ubiquitin-like domain. - Human spliceosome associated protein 114 (SAP 114 or SF3A120). - Yeast protein DSK2, a protein involved in spindle pole body duplication and which contains a N-terminal ubiquitin-like domain. - Human protein CKAP1/TFCB, *Schizosaccharomyces pombe* protein alp11 and *Caenorhabditis elegans* hypothetical protein F53F4.3. These proteins contain a N-terminal ubiquitin domain and a C-terminal CAP-Gly domain. - *Schizosaccharomyces pombe* hypothetical protein SpAC26A3.16. This protein contains a N-terminal ubiquitin domain. - Yeast protein SMT3. - Human ubiquitin-like proteins SMT3A and SMT3B. - Human ubiquitin-like protein SMT3C (also known as PIC1; Ubl1, Sumo-1; Gmp-1 or Sentrin). This protein is involved in targeting ranGAP1 to the nuclear pore complex protein ranBP2. - SMT3-like proteins in plants and *Caenorhabditis elegans*. To identify ubiquitin and related proteins, a pattern has been developed based on conserved positions in the central section of the sequence. A profile was also developed that spans the complete length of the ubiquitin domain.

Consensus pattern: K-x(2)-[LIVM]-x-[DESAK]-x(3)-[LIVM]-[PA]-x(3)-Q-x-[LIVM]-[LIVMC]-[LIVMFY]-x-G-x(4)-[DE]

[1] Jentsch S., Seufert W., Hauser H.-P. *Biochim. Biophys. Acta* 1089:127-139(1991).[2] Monia B.P., Ecker D.J., Croke S.T. *Bio/Technology* 8:209-215(1990).[3] Finley D., Varshavsky A. *Trends Biochem. Sci.* 10:343-347(1985).[4] Filippi M., Tribioli C., Toniolo D. *Genomics* 7:453-457(1990).[5] Olvera J., Wool I.G. *J. Biol. Chem.* 268:17967-17974(1993).[6] Kumar S., Yoshida Y., Noda M. *Biochem. Biophys. Res. Commun.* 195:393-399(1993).[7] Jones D., Candido E.P. *J. Biol. Chem.* 268:19545-19551(1993).[8] Melnick L., Sherman F. *J. Mol. Biol.* 233:372-388(1993).

Domain present in VPS-27, Hrs and STAM. Number of members: 27

711. Vinculin family signatures

Vinculin [1] is a eukaryotic protein that seems to be involved in the attachment of the actin-based microfilaments to the plasma membrane. Vinculin is located at the cytoplasmic side of focal contacts or adhesion plaques. In addition to actin, vinculin interacts with other structural proteins such as talin and alpha-actinins. Vinculin is a large protein of 116 Kd (about a 1000 residues). Structurally the protein consists of an acidic N-terminal domain of about 90 Kd separated from a basic C-terminal domain of about 25 Kd by a proline-rich region of about 50 residues. The central part of the N-terminal domain consists of a variable number (3 in vertebrates, 2 in *Caenorhabditis elegans*) of repeats of a 110 amino acids domain. Catenins [2] are proteins that associate with the cytoplasmic domain of a variety of cadherins. The association of catenins to cadherins produces a complex which is linked to the actin filament network, and which seems to be of primary importance for cadherins cell-adhesion properties. Three different types of catenins seem to exist: alpha, beta, and gamma. Alpha-catenins are proteins of about 100 Kd which are evolutionary related to vinculin. In terms of their structure the most significant differences are the absence, in alpha-catenin, of the repeated domain and of the proline-rich segment. Two signature patterns for this family of proteins have been developed. The first pattern is located in the N-terminal section of both vinculin and alpha-catenins and is part, in vinculin, of a domain that seems to be involved with the interaction with talin. The second pattern is based on a conserved region in the N-terminal part of the repeated domain of vinculin.

Consensus pattern: [KR]-x-[LIVMF]-x(3)-[LIVMA]-x(2)-[LIVM]-x(6)-R-Q-Q-E-L

Consensus pattern: [LIVM]-x-[QA]-A-x(2)-W-[IL]-x-[DN]-P

[1] Otto J.J. Cell Motil. Cytoskeleton 16:1-6(1990). [2] Herrenknecht K., Ozawa M., Eckerskorn C., Lottspeich F., Lenter M., Kemler R. Proc. Natl. Acad. Sci. U.S.A. 88:9156-9160(1991).

712. (Vitellogenin N) Lipoprotein amino terminal region

This family contains regions from: Vitellogenin, Microsomal triglyceride transfer protein and apolipoprotein B-100. These proteins are all involved in lipid transport [1]. This

570

family contains the LV1n chain from lipovitellin, that contains two structural domains.

Number of members: 33

[1] The structural basis of lipid interactions in lipovitellin, a soluble lipoprotein.

Anderson TA, Levitt DG, Banaszak LJ Structure 1998;6:895-909.

5

713. (VMSA) Major surface antigen from hepadnavirus

10 714. ssDNA binding protein (Viral DNA bp)

This protein is found in herpesviruses and is needed for replication.

15 715. (Votage CLC) Voltage gated chloride channels

This family of ion channels contains 10 or 12 transmembrane helices. Each protein forms a single pore. It has been shown that some members of this family form homodimers. These proteins contain two CBS domains.

[1] Schmidt-Rose T, Jentsch TJ; J Biol Chem 1997;272:20515-20521.

[2] Zhang J, George AL Jr, Griggs RC, Fouad GT, Roberts J, Kwiecinski H, Connolly AM, Ptacek LJ; Neurology 1996;47:993-998.

25

716. von Willebrand factor type A domain (vwa)

More von Willebrand factor type A domains? Sequence similarities with malaria thrombospondin-related anonymous protein, dihydropyridine-sensitive calcium channel and inter-alpha-trypsin inhibitor.

30

Bork P, Rohde K;

Biochem J 1991;279:908-911.

1. RUGGERI, Z.M. and WARE, J.
von Willebrand factor.
FASEB J. 7 308-316 (1993).

5 2. COLOMBATTI, A., BONALDO, P. and DOLIANA, R.
Type A modules: interacting domains found in several non-fibrillar
collagens and in other extracellular matrix proteins.
MATRIX 13 297-306 (1993).

10 3. PERKINS, S.J., SMITH, K.F., WILLIAMS, S.C., HARIS, P.I., CHAPMAN, D.
and SIM, R.B.
The secondary structure of the von Willebrand factor type A domain in
factor B of human complement by Fourier transform infrared spectroscopy.
Its occurrence in collagen types VI, VII, XII and XIV, the integrins and
other proteins by averaged structure predictions.
J.MOL.BIOL. 238 104-119 (1994).

4. BORK, P. and ROHDE, K.
More von Willebrand factor type A domains? Sequence similarities with
malaria thrombospondin-related anonymous protein, dihydropyridine-
sensitive calcium channel and inter-alpha-trypsin inhibitor.
BIOCHEM.J. 279 908-910 (1991).

5. EDWARDS, Y.J.K. and PERKINS, S.J.
25 The protein fold of the von Willebrand factor type A domain is predicted
to be similar to the open twisted beta-sheet flanked by alpha-helices
found in human ras-p21.
FEBS LETT. 358 283-286 (1995).

30 6. LEE, J.O., RIEU, P., ARNAOUT, M.A. and LIDDINGTON, R.
Crystal structure of the A domain from the alpha subunit of integrin CR3
(CD11b/CD18).
CELL 80 631-638 (1995).

7. QU, A. and LEAHY, D.J.

Crystal structure of the I-domain from the CD11a/CD18 (LFA-1, alpha L beta 2) integrin.

5 PROC.NATL.ACAD.SCI.USA 92 10277-10281 (1995).

The von Willebrand factor is a large multimeric glycoprotein found in blood plasma. Mutant forms are involved in the actiology of bleeding disorders [1]. In von Willebrand factor, the type A domain (vWF) is the prototype for a protein superfamily. The vWF domain is found in various plasma proteins: complement factors B, C2, CR3 and CR4; the integrins (I-domains); collagen types VI, VII, XII and XIV; and other extracellular proteins [2-4]. Proteins that incorporate vWF domains participate in numerous biological events (e.g., cell adhesion, migration, homing, pattern formation, and signal transduction), involving interaction with a large array of ligands [2]. Secondary structure prediction from 75 aligned vWF sequences has revealed a largely alternating sequence of alpha-helices and beta-strands [3]. Fold recognition algorithms were used to score sequence compatibility with a library of known structures: the vWF domain fold was predicted to be a doubly-wound, open, twisted beta-sheet flanked by alpha-helices [5]. 3D structures have been determined for the I-domains of integrins CD11b (with bound magnesium) [6] and CD11a (with bound manganese) [7]. The domain adopts a classic alpha/beta Rossmann fold and contains an unusual metal ion coordination site at its surface. It has been suggested that this site represents a general metal ion-dependent adhesion site (MIDAS) for binding protein ligands [6]. The residues constituting the MIDAS motif in the CD11b and CD11a I-domains are completely conserved, but the manner in which the metal ion is coordinated differs slightly [7].

30 VWFADOMAIN is a 3-element fingerprint that provides a signature for the vWF domain superfamily. The fingerprint was derived from an initial alignment of 14 sequences. Motif 1 includes the first beta-strand and 3 conserved residues involved in metal ion coordination in I-domains (Asp and 2 serines

573

in positions 8, 10 and 12, respectively); motif 2 spans strands beta-2 and beta-2'; and motif 3 encodes beta-strand 3 and a conserved Asp (in position 7), which coordinates the metal ion [6,7]. Three iterations on OWL27.0 were required to reach convergence, at which point a true set comprising 56 sequences was identified. Numerous partial matches were also found.

717. (WD40) WD domain, G-beta repeat

The ancient regulatory-protein family of WD-repeat proteins.

Neer EJ, Schmidt CJ, Nambudripad R, Smith TF;
Nature 1994;371:297-300.

Beta-transducin (G-beta) is one of the three subunits (alpha, beta, and gamma) of the guanine nucleotide-binding proteins (G proteins) which act as intermediaries in the transduction of signals generated by transmembrane receptors [1]. The alpha subunit binds to and hydrolyzes GTP; the functions of the beta and gamma subunits are less clear but they seem to be required for the replacement of GDP by GTP as well as for membrane anchoring and receptor recognition.

In higher eukaryotes G-beta exists as a small multigene family of highly conserved proteins of about 340 amino acid residues. Structurally G-beta consists of eight tandem repeats of about 40 residues, each containing a central Trp-Asp motif (this type of repeat is sometimes called a WD-40 repeat). Such a repetitive segment has been shown [E1,2,3,4,5] to exist in a number of other proteins listed below:

- Yeast STE4, a component of the pheromone response pathway. STE4 is a G-beta like protein that associates with GPA1 (G-alpha) and STE18 (G-gamma).
- Yeast MS11, a negative regulator of RAS-mediated cAMP synthesis. MS11 is most probably also a G-beta protein.
- Human and chicken protein 12.3. The function of this protein is not known, but on the basis of its similarity to G-beta proteins, it may also function

in signal transduction.

- *Chlamydomonas reinhardtii* gblp. This protein is most probably the homolog of vertebrate protein 12.3.

- Human LIS1, a neuronal protein involved in type-1 lissencephaly [E2].

5 - Mammalian coatamer beta' subunit (beta'-COP), a component of a cytosolic protein complex that reversibly associates with Golgi membranes to form vesicles that mediate biosynthetic protein transport.

10 - Yeast CDC4, essential for initiation of DNA replication and separation of the spindle pole bodies to form the poles of the mitotic spindle.

- Yeast CDC20, a protein required for two microtubule-dependent processes: nuclear movements prior to anaphase and chromosome separation.

- Yeast MAK11, essential for cell growth and for the replication of M1 double-stranded RNA.

15 - Yeast PRP4, a component of the U4/U6 small nuclear ribonucleoprotein with a probable role in mRNA splicing.

- Yeast PWP1, a protein of unknown function.

- Yeast SKI8, a protein essential for controlling the propagation of double-stranded RNA.

20 - Yeast SOF1, a protein required for ribosomal RNA processing which associates with U3 small nucleolar RNA.

- Yeast TUP1 (also known as AER2 or SFL2 or CYC9), a protein which has been implicated in dTMP uptake, catabolite repression, mating sterility, and many other phenotypes.

25 - Yeast YCR57c, an ORF of unknown function from chromosome III.

- Yeast YCR72c, an ORF of unknown function from chromosome III.

- Slime mold coronin, an actin-binding protein.

- Slime mold AAC3, a developmentally regulated protein of unknown function.

30

- *Drosophila* protein Groucho (formerly known as E(spl); 'enhancer of split'), a protein involved in neurogenesis and that seems to interact with the Notch and Delta proteins.

- Drosophila TAF-II-80, a protein that is tightly associated with TFIID.

The number of repeats in the above proteins varies between 5 (PRP4, TUP1, and Groucho) and 8 (G-beta, STE4, MSI1, AAC3, CDC4, PWP1, etc.). In G-beta and G-beta like proteins, the repeats span the entire length of the sequence, while in other proteins, they make up the N-terminal, the central or the C-terminal section.

A signature pattern can be developed from the central core of the domain (positions 9 to 23).

-Consensus pattern: [LIVMSTAC]-[LIVMFYWSTAGC]-[LIMSTAG]-[LIVMSTAGC]-x(2)-[DN]-x(2)-[LIVMWSTAC]-x-[LIVMFSTAG]-W-[DEN]-[LIVMFSTAGCN]

[1] Gilman A.G.

Annu. Rev. Biochem. 56:615-649(1987).

[2] Duronio R.J., Gordon J.I., Boguski M.S.

Proteins 13:41-56(1992).

[3] van der Voorn L., Ploegh H.L.

FEBS Lett. 307:131-134(1992).

[4] Neer E.J., Schmidt C.J., Nambudripad R., Smith T.F.

Nature 371:297-300(1994).

[5] Smith T.F., Gaiatzes C.G., Saxena K., Neer E.J.

Biochemistry In Press(1998).

718. WHEP-TRS domain containing proteins

A conserved domain of 46 amino acids has been shown [1] to exist in a number of higher eukaryote aminoacyl-transfer RNA synthetases. This domain is present one to six times in the following enzymes:

- Mammalian multifunctional aminoacyl-tRNA synthetase. The domain is present

576

three times in a region that separates the N-terminal glutamyl-tRNA synthetase domain from the C-terminal prolyl-tRNA synthetase domain.

- Drosophila multifunctional aminoacyl-tRNA synthetase. The domain is present six times in the intercatalytic region.

5 - Mammalian tryptophanyl-tRNA synthetase. The domain is found at the N-terminal extremity.

- Mammalian, insect, nematode and plant glycyl-tRNA synthetase. The domain is found at the N-terminal extremity [2].

10 - Mammalian histidyl-tRNA synthetase. The domain is found at the N-terminal extremity.

This domain, which is called WHEP-TRS, could contain a central alpha-helical region and may play a role in the association of tRNA-synthetases into multienzyme complexes.

5 A signature pattern based on the first 29 positions of the WHEP-Domain has been developed.

-Consensus pattern: [QY]-G-[DNEA]-x-[LIV]-[KR]-x(2)-K-x(2)-[KRNG]-[AS]-x(4)-
20 [LIV]-[DENK]-x(2)-[IV]-x(2)-L-x(3)-K

[1] Cerini C., Kerjan P., Astier M., Gratecos D., Mirande M., Semeriva M.
EMBO J. 10:4267-4277(1991).

[2] Nada S., Chang P.K., Dignam J.D.
25 J. Biol. Chem. 268:7660-7667(1993).

719. (Worm family 8) Putative membrane protein

Analysis of protein domain families in Caenorhabditis elegans.

30 Sonnhhammer EL, Durbin R;
Genomics 1997;46:200-216.

This family called family 8 in [1], may be a transmembrane protein

The specific function of this protein is unknown.

720. Xylose isomerase

Xylose isomerase (EC 5.3.1.5) [1] is an enzyme found in microorganisms which catalyzes the interconversion of D-xylose to D-xylulose. It can also isomerize D-ribose to D-ribulose and D-glucose to D-fructose. Xylose isomerase seems to require magnesium for its activity, while cobalt is necessary to stabilize the tetrameric structure of the enzyme. A number of residues are conserved in all known xylose isomerases.

Xylose isomerase also exists in plants [2] where it is homodimeric and is manganese-dependent.

Two signature patterns for xylose isomerase have been developed. The first one is derived from a stretch of five conserved amino acids that includes a glutamic acid residue known to be one of the four residues involved in the binding of the magnesium ion [3]; this pattern also includes a lysine residue which is involved in the catalytic activity. The second pattern is derived from a conserved region in the N-terminal section of the enzyme that includes a histidine residue which has been shown [4] to be involved in the catalytic mechanism of the enzyme.

-Consensus pattern: [LI]-E-P-K-P-x(2)-P

[E is a magnesium ligand]

[K is an active site residue]

-Consensus pattern: [FL]-H-D-x-D-[LIV]-x-[PD]-x-[GDE]

[H is an active site residue]

[1] Dauter Z., Dauter M., Hemker J., Witzel H., Wilson K.S.

FEBS Lett. 247:1-8(1989).

[2] Kristo P.A., Saarelainen R., Fagerstrom R., Aho S., Korhola M.

Eur. J. Biochem. 237:240-246(1996).

[3] Henrick K., Collyer C.A., Blow D.M.

J. Mol. Biol. 208:129-157(1989).

[4] Vangrysperre W., Ampe C., Kersters-Hilderson H., Tempst P.

Biochem. J. 263:195-199(1989).

5

721. XPG protein signatures. Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair. There are a minimum of seven genetic

10

complementation groups involved in this pathway: XP-A to XP-G. The defect in XP-G can be corrected by a 133 Kd nuclear protein called XPG (or XPGC) [2]. XPG belongs to a family of proteins [2,3,4,5,6] that are composed of two main subsets: - Subset 1, to which belongs XPG, RAD2 from budding yeast and rad13 from fission yeast. RAD2 and XPG are single-stranded DNA endonucleases [7,8]. XPG makes the 3' incision in human DNA nucleotide excision repair [9]. - Subset 2, to which belongs mouse and human FEN-1, rad2 from fission

15

yeast, and RAD27 from budding yeast. FEN-1 is a structure-specific endonuclease. In addition to the proteins listed in the above groups, this family also includes: - Fission yeast exo1, a 5'->3' double-stranded DNA exonuclease that could act in a pathway that corrects mismatched base pairs. - Yeast EXO1 (DHS1), a protein with probably the same function as exo1. - Yeast DIN7. Sequence alignment of this family of proteins reveals that similarities are largely confined to two regions. The first is located at the N-terminal extremity (N-region) and corresponds to the first 95 to 105 amino acids. The second region is internal (I-region) and found towards the C-terminus; it spans about 140 residues and contains a highly conserved core of 27 amino acids that includes a conserved pentapeptide (E-A-[DE]-A-[QS]).

20

It is possible that the conserved acidic residues are involved in the catalytic mechanism of DNA excision repair in XPG. The amino acids linking the N- and I-regions are not conserved; indeed, they are largely absent from proteins belonging to the second subset. Two signature patterns have been developed for these proteins. The first corresponds to the central part of the N-region, the second to part of the I-region and includes the putative catalytic core pentapeptide

30

Consensus pattern: [VI]-[KRE]-P-x-[FYIL]-V-F-D-G-x(2)-[PIL]-x-[LVC]-K-

Consensus pattern: [GS]-[LIVM]-[PER]-[FYS]-[LIVM]-x-A-P-x-E-A-[DE]-[PAS]- [QS]-
[CLM]-

[1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).[2] Scherly D., Nospikel
5 T., Corlet J., Ucla C., Bairoch A., Clarkson S.G. Nature 363:182-185(1993).[3] Carr A.M.,
Sheldrick K.S., Murray J.M., Al-Harithy R., Watts F.Z., Lehmann A.R. Nucleic Acids Res.
21:1345-1349(1993).[4] Murray J.M., Tavassoli M., Al-Harithy R., Sheldrick K.S.,
Lehmann A.R., Carr A.M., Watts F.Z. Mol. Cell. Biol. 14:4878-4888(1994).[5] Harrington
10 J.J., Lieber M.R. Genes Dev. 8:1344-1355(1994).[6] Szankasi P., Smith G.R. Science
267:1166-1169(1995).[7] Habraken Y., Sung P., Prakash L., Prakash S. Nature 366:365-
368(1993).[8] O'Donovan A., Scherly D., Clarkson S.G., Wood R.D. J. Biol. Chem.
269:15965-15968(1994).[9] O'Donovan A., Davies A.A., Moggs J.G., West S.C., Wood
R.D. Nature 371:432-435(1994).

722. Xanthine/uracil permeases family

The following transport proteins which are involved in the uptake of xanthine
or uracil are evolutionary related [1]:

- Uric acid-xanthine permease (gene uapA) from *Aspergillus nidulans*.
- Purine permease (gene uapC) from *Aspergillus nidulans*.
- Xanthine permease from *Bacillus subtilis* (gene pbuX).
- Uracil permease from *Escherichia coli* (gene uraA) [2] and *Bacillus* (gene
pyrP).
- Hypothetical protein ycdG from *Escherichia coli*.
- Hypothetical protein ygfO from *Escherichia coli*.
- Hypothetical protein ygfU from *Escherichia coli*.
- Hypothetical protein yicE from *Escherichia coli*.
- Hypothetical protein yunJ from *Bacillus subtilis*.
- Hypothetical protein yunK from *Bacillus subtilis*.

They are proteins of from 430 to 595 residues that seem to contain 12
transmembrane domains.

The best conserved region which corresponds with what seems to be the tenth transmembrane domain has been selected as a signature pattern.

-Consensus pattern: [LIVM]-P-x-[PASIF]-V-[LIVM]-G-G-x(4)-[LIVM]-[FY]-[GSA]-x-[LIVM]-x(3)-G

[1] Diallinas G., Gorfinkiel L., Arst G., Cecchetto G., Scazzocchio C.

J. Biol. Chem. 270:8610-8622(1995).

[2] Andersen P.S., Frees D., Fast R., Mygind B.

J. Bacteriol. 177:2008-2013(1995).

723. Hypothetical yabO/yceC/sfhB family

The following proteins, which seems to belong to a family of pseudouridine synthases (EC 4.2.1.70) [1] have been shown to share regions of similarities:

- Escherichia coli and Haemophilus influenzae ribosomal large subunit pseudouridine synthase A (gene rluA). It is responsible for synthesis of pseudouridine from uracil-746 IN 23S rRNA.
- Escherichia coli and Haemophilus influenzae ribosomal large subunit pseudouridine synthase C (gene rluC). It is responsible for synthesis of pseudouridine from uracil at positions 955, 2504 and 2580 in 23S rRNA.
- Escherichia coli protein and homologs in other bacteria large subunit pseudouridine synthase D (gene rluD).
- Yeast DRAP deaminase (gene RIB2).
- Escherichia coli hypothetical protein yqcB and HI1435, the corresponding Haemophilus influenzae protein.
- Haemophilus influenzae hypothetical protein HI0042.
- Aquifex aeolicus hypothetical protein AQ_1758.
- Bacillus subtilis hypothetical protein yhcT.
- Bacillus subtilis hypothetical protein yjbO.
- Bacillus subtilis hypothetical protein ylyB.
- Helicobacter pylori hypothetical protein HP0347.
- Helicobacter pylori hypothetical protein HP0745.

581

- Helicobacter pylori hypothetical protein HP0956.
- Mycoplasma genitalium hypothetical protein MG209.
- Mycoplasma genitalium hypothetical protein MG370.
- Synechocystis strain PCC 6803 hypothetical protein slr1592.
- 5 - Synechocystis strain PCC 6803 hypothetical protein slr1629.
- Yeast hypothetical protein YDL036c.
- Yeast hypothetical protein YGR169c.
- Fission yeast hypothetical protein SpAC18B11.02c.
- Caenorhabditis elegans hypothetical protein K07E8.7.

10

These are proteins of from 21 to 50 Kd which contain a number of conserved regions in their central section. They can be picked up in the database by the following highly conserved pattern.

15

-Consensus pattern: [LIVCA]-[NHYT]-R-[LI]-D-x(2)-T-[STA]-G-[LIVAGC]-
[LIVMF](2)-[LIVMFGC]-[SGTACV]

20

[1] Conrad J., Sun D., Englund N., Ofengand J.
J. Biol. Chem. 273:18562-18566(1998).

In addition, the following bacterial proteins, which seems to belong to a family of pseudouridine synthases (EC 4.2.1.70) [1] also have been shown to share regions of similarities:

25

- Escherichia coli and Haemophilus influenzae 16S pseudouridylate 516 synthase (EC 4.2.1.70) (gene: rsuA). This enzyme is responsible for the formation of pseudouridine from uracil-516 in 16S ribosomal RNA.
- Escherichia coli hypothetical protein yciL and HI1199, the corresponding Haemophilus influenzae protein.

30

- Escherichia coli hypothetical protein yjbC.
- Escherichia coli hypothetical protein ymfC and HI0694, the corresponding Haemophilus influenzae protein.
- Aquifex aeolicus hypothetical protein AQ_554.

582

- Aquifex aeolicus hypothetical protein AQ_1464.
- Bacillus subtilis hypothetical protein ypuL.
- Bacillus subtilis hypothetical protein ytzF.
- Borrelia burgdorferi hypothetical protein BB0129.
- 5 - Helicobacter pylori hypothetical protein HP1459.
- Synechocystis strain PCC 6803 hypothetical protein slr0361.
- Synechocystis strain PCC 6803 hypothetical protein slr0612.

These are proteins of from 25 to 40 Kd which contain a number of conserved
 10 regions in their central section. They can be picked up in the database by the
 following highly conserved pattern.

-Consensus pattern: G-R-L-D-x(2)-[STA]-x-G-[LIVFA]-[LIVMF](3)-[ST]-[DNST]

5 [1] Wrzesinski J., Bakin A., Nurse K., Lane B.G., Ofengand J.
 Biochemistry 34:8904-8913(1995).

724. Zinc finger present in dystrophin, CBP/p300

20 ZZ in dystrophin binds calmodulin

Putative zinc finger; binding not yet shown.

725..Zinc carboxypeptidase

25 There are a number of different types of zinc-dependent carboxypeptidases (EC
 3.4.17.-) [1,2]. All these enzymes seem to be structurally and functionally
 related. The enzymes that belong to this family are listed below.

- Carboxypeptidase A1 (EC 3.4.17.1), a pancreatic digestive enzyme that can
 30 removes all C-terminal amino acids with the exception of Arg, Lys and Pro.
- Carboxypeptidase A2 (EC 3.4.17.15), a pancreatic digestive enzyme with a
 specificity similar to that of carboxypeptidase A1, but with a preference
 for bulkier C-terminal residues.

- Carboxypeptidase B (EC 3.4.17.2), also a pancreatic digestive enzyme, but that preferentially removes C-terminal Arg and Lys.
- Carboxypeptidase N (EC 3.4.17.3) (also known as arginine carboxypeptidase), a plasma enzyme which protects the body from potent vasoactive and inflammatory peptides containing C-terminal Arg or Lys (such as kinins or anaphylatoxins) which are released into the circulation.
- Carboxypeptidase H (EC 3.4.17.10) (also known as enkephalin convertase or carboxypeptidase E), an enzyme located in secretory granules of pancreatic islets, adrenal gland, pituitary and brain. This enzyme removes residual C-terminal Arg or Lys remaining after initial endoprotease cleavage during prohormone processing.
- Carboxypeptidase M (EC 3.4.17.12), a membrane bound Arg and Lys specific enzyme.
It is ideally situated to act on peptide hormones at local tissue sites where it could control their activity before or after interaction with specific plasma membrane receptors.
- Mast cell carboxypeptidase (EC 3.4.17.1), an enzyme with a specificity to carboxypeptidase A, but found in the secretory granules of mast cells.
- *Streptomyces griseus* carboxypeptidase (Cpase SG) (EC 3.4.17.-) [3], which combines the specificities of mammalian carboxypeptidases A and B.
- *Thermoactinomyces vulgaris* carboxypeptidase T (EC 3.4.17.18) (CPT) [4], which also combines the specificities of carboxypeptidases A and B.
- AEBP1 [5], a transcriptional repressor active in preadipocytes. AEBP1 seems to regulate transcription by cleavage of other transcriptional proteins.
- Yeast hypothetical protein YHR132c.

All of these enzymes bind an atom of zinc. Three conserved residues are implicated in the binding of the zinc atom: two histidines and a glutamic acid. Two signature patterns which contain these three zinc-ligands have been derived.

-Consensus pattern: [PK]-x-[LIVMFY]-x-[LIVMFY]-x(4)-H-[STAG]-x-E-x-[LIVM]-[STAG]-x(6)-[LIVMFYTA]
[H and E are zinc ligands]

-Consensus pattern: H-[STAG]-x(3)-[LIVME]-x(2)-[LIVMFYW]-P-[FYW]
[H is a zinc ligand]

[1] Tan F., Chan S.J., Steiner D.F., Schilling J.W., Skidgel R.A.

5 J. Biol. Chem. 264:13165-13170(1989).

[2] Reynolds D.S., Stevens R.L., Gurley D.S., Lane W.S., Austen K.F.,
Serafin W.E.

J. Biol. Chem. 264:20094-20099(1989).

[3] Narahashi Y.

10 J. Biochem. 107:879-886(1990).

[4] Teplyakov A., Polyakov K., Obmolova G., Strokopytov B., Kuranova I.,
Osterman A.L., Grishin N.V., Smulevitch S.V., Zagnitko O.P.,
Galperina O.V., Matz M.V., Stepanov V.M.
Eur. J. Biochem. 208:281-288(1992).

15 [5] He G.-P., Muise A., Li A.W., Ro H.-S.
Nature 378:92-96(1995).

[6] Hourdou M.-L., Guinand M., Vacheron M.J., Michel G., Denoroy L.,
Duez C.M., Englebert S., Joris B., Weber G., Ghuysen J.-M.
Biochem. J. 292:563-570(1993).

20 [7] Rawlings N.D., Barrett A.J.
Meth. Enzymol. 248:183-228(1995).

726. Zinc finger, C2H2 type

25 The C2H2 zinc finger is the classical zinc finger domain.

The two conserved cysteines and histidines co-ordinate a
zinc ion. The following pattern describes the zinc finger.

#-X-C-X(1-5)-C-X3-#-X5-#-X2-H-X(3-6)-[H/C]

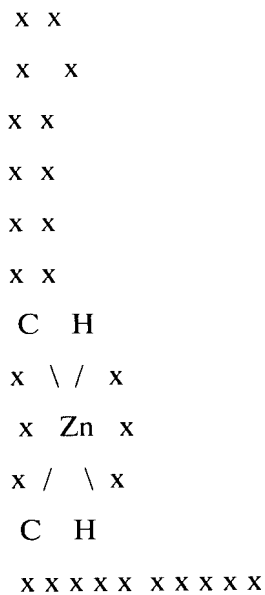
Where X can be any amino acid, and numbers in brackets

30 indicate the number of residues. The positions marked # are
those that are important for the stable fold of the zinc
finger. The final position can be either his or cys.

The C2H2 zinc finger is composed of two short beta strands

followed by an alpha helix. The amino terminal part of the helix binds the major groove in DNA binding zinc fingers.

'Zinc finger' domains [1-5] are nucleic acid-binding protein structures first identified in the *Xenopus* transcription factor TFIIIA. These domains have since been found in numerous nucleic acid-binding proteins. A zinc finger domain is composed of 25 to 30 amino-acid residues. There are two cysteine or histidine residues at both extremities of the domain, which are involved in the tetrahedral coordination of a zinc atom. It has been proposed that such a domain interacts with about five nucleotides. A schematic representation of a zinc finger domain is shown below:



Many classes of zinc fingers are characterized according to the number and positions of the histidine and cysteine residues involved in the zinc atom coordination. In the first class to be characterized, called C2H2, the first pair of zinc coordinating residues are cysteines, while the second pair are histidines. A number of experimental reports have demonstrated the zinc-dependent DNA or RNA binding property of some members of this class.

Some of the proteins known to include C2H2-type zinc fingers are listed below.

The number of zinc finger regions found in each of these proteins are indicated between brackets; a '+' symbol indicates that only partial sequence data is available and that additional finger domains may be present.

- 5 - *Saccharomyces cerevisiae*: ACE2 (3), ADR1 (2), AZF1 (4), FZF1 (5), MIG1 (2), MSN2 (2), MSN4 (2), RGM1 (2), RIM1 (3), RME1 (3), SFP1 (2), SSL1 (1), STP1 (3), SWI5 (3), VAC1 (1) and ZMS1 (2).
- *Emmericella nidulans*: brlA (2), creA (2).
- *Drosophila*: AEF-1 (4), Cf2 (7), ci-D (5), Disconnected (2), Escargot (5),
 10 Glass (5), Hunchback (6), Kruppel (5), Kruppel-H (4+), Odd-skipped (4), Odd-paired (4), Pep (3), Snail (5), Spalt-major (7), Serependity locus beta (6), delta (7), h-1 (8), Suppressor of hairy wing su(Hw) (12), Suppressor of variegation suvar(3)7 (5), Teashirt (3) and Tramtrack (2).
- *Xenopus*: transcription factor TFIIIA (9), p43 from RNP particle (9), Xfin (37 !!), Xsna (5), gastrula XlcGF5.1 to XlcGF71.1 (from 4+ to 11+), Oocyte XlcOF2 to XlcOF22 (from 7 to 12).
- Mammalian: basonuclin (6), BCL-6/LAZ-3 (6), erythroid krueppel-like transcription factor (3), transcription factors Sp1 (3), Sp2 (3), Sp3 (3) and Sp(4) 3, transcriptional repressor YY1 (4), Wilms' tumor protein (4),
 15 EGR1/Krox24 (3), EGR2/Krox20 (3), EGR3/Pilot (3), EGR4/AT133 (4), Evi-1 (10), GLI1 (5), GLI2 (4+), GLI3 (3+), HIV-EP1/ZNF40 (4), HIV-EP2 (2), KR1 (9+), KR2 (9), KR3 (15+), KR4 (14+), KR5 (11+), HF.12 (6+), REX-1 (4), ZfX (13), ZfY (13), Zfp-35 (18), ZNF7 (15), ZNF8 (7), ZNF35 (10), ZNF42/MZF-1 (13), ZNF43 (22), ZNF46/Kup (2), ZNF76 (7), ZNF91 (36), ZNF133 (3).

25

In addition to the conserved zinc ligand residues it has been shown [6] that a number of other positions are also important for the structural integrity of the C2H2 zinc fingers. The best conserved position is found four residues after the second cysteine; it is generally an aromatic or aliphatic residue.

30

-Consensus pattern: C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H
 [The two C's and two H's are zinc ligands]

- [1] Klug A., Rhodes D.
Trends Biochem. Sci. 12:464-469(1987).
- [2] Evans R.M., Hollenberg S.M.
Cell 52:1-3(1988).
- 5 [3] Payre F., Vincent A.
FEBS Lett. 234:245-250(1988).
- [4] Miller J., McLachlan A.D., Klug A.
EMBO J. 4:1609-1614(1985).
- [5] Berg J.M.
10 Proc. Natl. Acad. Sci. U.S.A. 85:99-102(1988).
- [6] Rosenfeld R., Margalit H.
J. Biomol. Struct. Dyn. 11:557-570(1993).

15 727. Zinc finger, C3HC4 type (RING finger)

A number of eukaryotic and viral proteins contain a conserved cysteine-rich domain of 40 to 60 residues (called C3HC4 zinc-finger or 'RING' finger) [1] that binds two atoms of zinc, and is probably involved in mediating protein-protein interactions. The 3D structure of the zinc ligation system is unique to the RING domain and is referred to as the "cross-brace" motif. The spacing of the cysteines in such a domain is C-x(2)-C-x(9 to 39)-C-x(1 to 3)-H-x(2 to 3)-C-x(2)-C-x(4 to 48)-C-x(2)-C.

20
25 Proteins currently known to include the C3HC4 domain are listed below (references are only provided for recently determined sequences).

- Mammalian V(D)J recombination activating protein (gene RAG1). RAG1 activates the rearrangement of immunoglobulin and T-cell receptor genes.
- Mouse rpt-1. Rpt-1 is a trans-acting factor that regulates gene expression
30 directed by the promoter region of the interleukin-2 receptor alpha chain or the LTR promoter region of HIV-1.
- Human rfp. Rfp is a developmentally regulated protein that may function in male germ cell development. Recombination of the N-terminal section of rfp

with a protein tyrosine kinase produces the ret transforming protein.

- Human 52 Kd Ro/SS-A protein. A protein of unknown function from the Ro/SS-A ribonucleoprotein complex. Sera from patients with systemic lupus erythematosus or primary Sjogren's syndrome often contain antibodies that react with the Ro proteins.

- Human histocompatibility locus protein RING1.

- Human PML, a probable transcription factor. Chromosomal translocation of PML with retinoic receptor alpha creates a fusion protein which is the cause of acute promyelocytic leukemia (APL).

- Mammalian breast cancer type 1 susceptibility protein (BRCA1) [E1].

- Mammalian cbl proto-oncogene.

- Mammalian bmi-1 proto-oncogene.

- Vertebrate CDK-activating kinase (CAK) assembly factor MAT1, a protein that stabilizes the complex between the CDK7 kinase and cyclin H (MAT1 stands for 'Menage A Trois').

- Mammalian mel-18 protein. Mel-18 which is expressed in a variety of tumor cells is a transcriptional repressor that recognizes and bind a specific DNA sequence.

- Mammalian peroxisome assembly factor-1 (PAF-1) (PMP35), which is somewhat involved in the biogenesis of peroxisomes. In humans, defects in PAF-1 are responsible for a form of Zellweger syndrome, an autosomal recessive disorder associated with peroxisomal deficiencies.

- Human MAT1 protein, which interacts with the CDK7-cyclin H complex.

- Human RING1 protein.

- Xenopus XNF7 protein, a probable transcription factor.

- Trypanosoma protein ESAG-8 (T-LR), which may be involved in the postranscriptional regulation of genes in VSG expression sites or may interact with adenylate cyclase to regulate its activity.

- Drosophila proteins Posterior Sex Combs (Psc) and Suppressor two of zeste (Su(z)2). The two proteins belong to the Polycomb group of genes needed to maintain the segment-specific repression of homeotic selector genes.

- Drosophila protein male-specific msl-2, a DNA-binding protein which is involved in X chromosome dosage compensation (the elevation of

589

transcription of the male single X chromosome).

- Arabidopsis thaliana protein COP1 which is involved in the regulation of photomorphogenesis.

- Fungal DNA repair proteins RAD5, RAD16, RAD18 and rad8.

5 - Herpesviruses trans-acting transcriptional protein ICP0/IE110. This protein which has been characterized in many different herpesviruses is a trans-activator and/or -repressor of the expression of many viral and cellular promoters.

- Baculoviruses protein CG30.

10 - Baculoviruses major immediate early protein (PE-38).

- Baculoviruses immediate-early regulatory protein IE-N/IE-2.

- Caenorhabditis elegans hypothetical proteins F54G8.4, R05D3.4 and T02C1.1.

- Yeast hypothetical proteins YER116c and YKR017c.

15 The central region of the domain was selected as a signature pattern for the C3HC4 finger.

-Consensus pattern: C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMYA]

20 [1] Borden K.L.B., Freemont P.S.

Curr. Opin. Struct. Biol. 6:395-401(1996).

728. Zinc finger C-x8-C-x5-C-x3-H type (and similar).

729. Zinc finger, CCHC class

A family of CCHC zinc fingers, mostly from retroviral gag proteins (nucleocapsid). Prototype structure is from HIV.

30 Also contains members involved in eukaryotic gene regulation, such as C. elegans GLH-1.

Structure is an 18-residue zinc finger; no examples of indels in the alignment.

730. Zn-finger in Ran binding protein and others.

5

731. AN1-like Zinc finger

Zinc finger at the C-terminus of An1 Swiss:Q91889, a ubiquitin-like protein in *Xenopus laevis*. The following pattern describes the zinc finger. C-X2-C-X(9-12)-C-X(1-2)-C-X4-C-
 10 X2-H-X5-H-X-C Where X can be any amino acid, and numbers in brackets indicate the number of residues.

[1] Linnen JM, Bailey CP, Weeks DL; Gene 1993;128:181-188.

732. 14-3-3 proteins

Structure of a 14-3-3 protein and implications for coordination of multiple signalling pathways.

Xiao B, Smerdon SJ, Jones DH, Dodson GG, Soneji Y, Aitken A, Gamblin SJ;
 20 Nature 1995;376:188-191.

Crystal structure of the zeta isoform of the 14-3-3 protein.

Liu D, Bienkowska J, Petosa C, Collier RJ, Fu H, Liddington R;
 Nature 1995;376:191-194.

25 Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine.

Muslin AJ, Tanner JW, Allen PM, Shaw AS;
 Cell 1996;84:889-897.

30 The 14-3-3 protein binds its target proteins with a common site located towards the C-terminus.

Ichimura T, Ito M, Itagaki C, Takahashi M, Horigome T, Omata S, Ohno S,
 Isobe T

FEBS Lett 1997;413:273-276.

Molecular evolution of the 14-3-3 protein family.

Wang W, Shakes DC

5 J Mol Evol 1996;43:384-398.

Function of 14-3-3 proteins.

Jin DY, Lyu MS, Kozak CA, Jeang KT

Nature 1996;382:308-308.

10 The 14-3-3 proteins [1,2,3] are a family of closely related acidic homodimeric proteins of about 30 Kd which were first identified as being very abundant in mammalian brain tissues and located preferentially in neurons. The 14-3-3 proteins seem to have multiple biological activities and play a key role in signal transduction pathways and the cell cycle. They interact with kinases such as PKC or Raf-1; they seem to also function as protein-kinase dependent
15 activators of tyrosine and tryptophan hydroxylases and in plants they are associated with a complex that binds to the G-box promoter elements.

The 14-3-3 family of proteins are ubiquitously found in all eukaryotic species
20 studied and have been sequenced in fungi (yeast BMH1 and BMH2, fission yeast rad24 and rad25), plants, Drosophila, and vertebrates. The sequences of the 14-3-3 proteins are extremely well conserved. Two highly conserved regions have been selected as signature patterns: the first is a peptide of 11 residues located in the N-terminal section; the second, a 20 amino acid region located
25 in the C-terminal section.

-Consensus pattern: R-N-L-[LIV]-S-[VG]-[GA]-Y-[KN]-N-[IVA]

-Consensus pattern: Y-K-[DE]-S-T-L-I-[IM]-Q-L-[LF]-[RHC]-D-N-[LF]-T-[LS]-W-[TAN]-[SAD]

30

[1] Aitken A.

Trends Biochem. Sci. 20:95-97(1995).

[2] Morrison D.

Science 266:56-57(1994).

- [3] Xiao B., Smerdon S.J., Jones D.H., Dodson G.G., Soneji Y., Aitken A.,
Gamblin S.J.
Nature 376:188-191(1995).

5

733. D-isomer specific 2-hydroxyacid dehydrogenases (2 Hacid DH)

This Pfam covers the Formate dehydrogenase, D-glycerate dehydrogenase and D-lactate dehydrogenase families in SCOP. A number of NAD-dependent 2-hydroxyacid dehydrogenases which seem to be specific for the D-isomer of their substrate have been shown [1,2,3,4] to be functionally and structurally related. These enzymes are listed below.

10

- D-lactate dehydrogenase (EC 1.1.1.28), a bacterial enzyme which catalyzes the reduction of D-lactate to pyruvate.
- D-glycerate dehydrogenase (EC 1.1.1.29) (NADH-dependent hydroxypyruvate reductase), a plant leaf peroxisomal enzyme that catalyzes the reduction of hydroxypyruvate to glycerate. This reaction is part of the glycolate pathway of photorespiration.
- D-glycerate dehydrogenase from the bacteria *Hyphomicrobium methylovorum* and *Methylobacterium extorquens*.
- 3-phosphoglycerate dehydrogenase (EC 1.1.1.95), a bacterial enzyme that catalyzes the oxidation of D-3-phosphoglycerate to 3-phosphohydroxypyruvate. This reaction is the first committed step in the 'phosphorylated' pathway of serine biosynthesis.
- Erythronate-4-phosphate dehydrogenase (EC 1.1.1.-) (gene *pdxB*), a bacterial enzyme involved in the biosynthesis of pyridoxine (vitamin B6).
- D-2-hydroxyisocaproate dehydrogenase (EC 1.1.1.-) (D-hicDH), a bacterial enzyme that catalyzes the reversible and stereospecific interconversion between 2-ketocarboxylic acids and D-2-hydroxy-carboxylic acids.
- Formate dehydrogenase (EC 1.2.1.2) (FDH) from the bacteria *Pseudomonas* sp. 101 and various fungi [5].
- Vancomycin resistance protein *vanH* from *Enterococcus faecium*; this protein is a D-specific alpha-keto acid dehydrogenase involved in the formation of a peptidoglycan which does not terminate by D-alanine thus preventing vancomycin binding.
- *Escherichia coli* hypothetical protein *ycdW*.

15

20

25

30

35

593

- Escherichia coli hypothetical protein yiaE.
- Haemophilus influenzae hypothetical protein HI1556.
- Yeast hypothetical protein YER081w.
- Yeast hypothetical protein YIL074w.

5 All these enzymes have similar enzymatic activities and are structurally related. Three of the most conserved regions of these proteins have been selected to develop patterns. The first pattern is based on a glycine-rich region located in the central section of these enzymes; this region probably corresponds to the NAD-binding domain. The two other patterns contain a number of conserved charged residues, some of which may play a role in the catalytic
10 mechanism.

-Consensus pattern: [LIVMA]-[AG]-[IVT]-[LIVMFY]-[AG]-x-G-[NHKROGSAC]-[LIV]-G-x(13,14)-[LIVfMT]-x(2)-[FYwCTH]-[DNSTK]

-Consensus pattern: [LIVMFYWA]-[LIVFYWC]-x(2)-[SAC]-[DNQHR]-[IVFA]-[LIVF]-x-[LIVF]-[HNI]-x-P-x(4)-[STN]-x(2)-[LIVMF]-x-[GSDN]

-Consensus pattern: [LMFATC]-[KPQ]-x-[GSTDN]-x-[LIVMFYWR]-[LIVMFYW](2)-N-x-[STAGC]-R-[GP]-x-[LIVH]-[LIVMC]-[DNV]

[1] Grant G.A. Biochem. Biophys. Res. Commun. 165:1371-1374(1989).

[2] Kochhar S., Hunziker P., Leong-Morgenthaler P.M., Hottinger H. Biochem. Biophys. Res. Commun. 184:60-66(1992).

[3] Ohta T., Taguchi H. J. Biol. Chem. 266:12588-12594(1991).

[4] Goldberg J.D., Yoshida T., Brick P. J. Mol. Biol. 236:1123-1140(1994).

[5] Popov V.O., Lamzin V.S. Biochem. J. 301:625-643(1994).

734. 2-oxo acid dehydrogenases acyltransferase (catalytic domain)

Refined crystal structure of the catalytic domain of dihydrolipoyl transacetylase (E2P) from azotobacter vineelandii at 2.6 angstroms
30 resolution.

Mattevi A, Obmolova G, Kalk KH, Westphal AH, De Kok A, Hol WG;
J Mol Biol 1993;230:1183-1199.

These proteins contain one to three copies of a lipoyl binding domain

followed by the catalytic domain.

735. 3-beta hydroxysteroid dehydrogenase/isomerase family

- 5 Structure and tissue-specific expression of 3
beta-hydroxysteroid dehydrogenase/5-ene-4-ene isomerase
genes in human and rat classical and peripheral
steroidogenic tissues.

Labrie F, Simard J, Luu-The V, Pelletier G, Belanger A,

- 10 Lachance Y, Zhao HF, Labrie C, Breton N, de Launoit Y, et al
J Steroid Biochem Mol Biol 1992;41:421-435.

The enzyme 3 beta-hydroxysteroid dehydrogenase/5-ene-4-ene
isomerase (3 beta-HSD) catalyzes the oxidation and isomerization
of 5-ene-3 beta-hydroxypregnene and 5-ene-hydroxyandrostene
15 steroid precursors into the corresponding 4-ene-ketosteroids necessary
for the formation of all classes of steroid hormones.

736. 3-hydroxyacyl-CoA dehydrogenase

20 This family also includes lambda crystallin.

Structure of L-3-hydroxyacyl-coenzyme A dehydrogenase:
preliminary chain tracing at 2.8-A resolution.

Birktoft JJ, Holden HM, Hamlin R, Xuong NH, Banaszak LJ;
Proc Natl Acad Sci U S A 1987;84:8262-8266.

- 25 3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35) (HCDH) [1] is an enzyme involved
in fatty acid metabolism, it catalyzes the reduction of 3-hydroxyacyl-CoA to
3-oxoacyl-CoA. Most eukaryotic cells have 2 fatty-acid beta-oxidation systems,
one located in mitochondria and the other in peroxisomes. In peroxisomes
30 3-hydroxyacyl-CoA dehydrogenase forms, with enoyl-CoA hydratase (ECH) and
3,2-trans-enoyl-CoA isomerase (ECI) a multifunctional enzyme where the N-
terminal domain bears the hydratase/isomerase activities and the C-terminal
domain the dehydrogenase activity. There are two mitochondrial enzymes: one

which is monofunctional and the other which is, like its peroxisomal counterpart, multifunctional.

In *Escherichia coli* (gene *fadB*) and *Pseudomonas fragi* (gene *faoA*) HCDH is part of a multifunctional enzyme which also contains an ECH/ECI domain as well as a 3-hydroxybutyryl-CoA epimerase domain [2].

The other proteins structurally related to HCDH are:

- Bacterial 3-hydroxybutyryl-CoA dehydrogenase (EC 1.1.1.157) which reduces 3-hydroxybutanoyl-CoA to acetoacetyl-CoA [3].
- Eye lens protein lambda-crystallin [4], which is specific to lagomorphes (such as rabbit).

There are two major region of similarities in the sequences of proteins of the HCDH family, the first one located in the N-terminal, corresponds to the NAD-binding site, the second one is located in the center of the sequence. A signature pattern has been derived from this central region.

-Consensus pattern: [DNE]-x(2)-[GA]-F-[LIVMFY]-x-[NT]-R-x(3)-[PA]-[LIVMFY](2)-x(5)-[LIVMFYCT]-[LIVMFY]-x(2)-[GV]

[1] Birktoff J.J., Holden H.M., Hamlin R., Xuong N.-H., Banaszak L.J.
Proc. Natl. Acad. Sci. U.S.A. 84:8262-8266(1987).

[2] Nakahigashi K., Inokuchi H.
Nucleic Acids Res. 18:4937-4937(1990).

[3] Mullany P., Clayton C.L., Pallen M.J., Slone R., Al-Saleh A.,
Tabaqchali S.
FEMS Microbiol. Lett. 124:61-67(1994).

[4] Mulders J.W.M., Hendriks W., Blankesteyn W.M., Bloemendal H.,
de Jong W.W.
J. Biol. Chem. 263:15462-15466(1988).

737. 60s Acidic ribosomal protein

Proteins P1, P2, and P0, components of the eukaryotic ribosome stalk. New structural and functional aspects.

- 5 Remacha M, Jimenez-Diaz A, Santos C, Briones E, Zambrano R, Rodriguez Gabriel MA, Guarinos E, Ballesta JP; Biochem Cell Biol 1995;73:959-968.

This family includes archaebacterial L12, eukaryotic P0, P1 and P2.

10

738. 6-phosphogluconate dehydrogenases

6-phosphogluconate dehydrogenase (EC 1.1.1.44) (6PGD) catalyzes the third step in the hexose monophosphate shunt, the decarboxylating reduction of 6-phosphogluconate in to ribulose 5-phosphate.

5

Prokaryotic and eukaryotic 6PGD are proteins of about 470 amino acids whose sequence are highly conserved [1]. A region which has been shown [2], from studies of the sheep 6PGD tertiary structure, to be involved in the binding of 6-phosphogluconate has been selected as a signature pattern.

20

-Consensus pattern: [LIVM]-x-D-x(2)-[GA]-[NQS]-K-G-T-G-x-W

[1] Reizer A., Deutscher J., Saier M.H. Jr., Reizer J. Mol. Microbiol. 5:1081-1089(1991).

- 25 [2] Adams M.J., Archibald I.G., Bugg C.E., Carne A., Gover S., Helliwell J.R., Pickersgill R.W., White S.W. EMBO J. 2:1009-1014(1983).

- 30 739. (7tm 1) G-protein coupled receptors [1 to 4,E1,E2] (also called R7G) are an extensive group of hormones, neurotransmitters, odorants and light receptors which transduce extracellular signals by interaction with guanine nucleotide-binding (G) proteins. The receptors that are currently known to belong to this

family are listed below.

- 5-hydroxytryptamine (serotonin) 1A to 1F, 2A to 2C, 4, 5A, 5B, 6 and 7 [5].
- Acetylcholine, muscarinic-type, M1 to M5.
- 5 - Adenosine A1, A2A, A2B and A3 [6].
- Adrenergic alpha-1A to -1C; alpha-2A to -2D; beta-1 to -3 [7].
- Angiotensin II types I and II.
- Bombesin subtypes 3 and 4.
- Bradykinin B1 and B2.
- 10 - c3a and C5a anaphylatoxin.
- Cannabinoid CB1 and CB2.
- Chemokines C-C CC-CKR-1 to CC-CKR-8.
- Chemokines C-X-C CXC-CKR-1 to CXC-CKR-4.
- Cholecystokinin-A and cholecystokinin-B/gastrin.
- 15 - Dopamine D1 to D5 [8].
- Endothelin ET-a and ET-b [9].
- fMet-Leu-Phe (fMLP) (N-formyl peptide).
- Follicle stimulating hormone (FSH-R) [10].
- Galanin.
- 20 - Gastrin-releasing peptide (GRP-R).
- Gonadotropin-releasing hormone (GNRH-R).
- Histamine H1 and H2 (gastric receptor I).
- Lutropin-choriogonadotropic hormone (LSH-R) [10].
- Melanocortin MC1R to MC5R.
- 25 - Melatonin.
- Neuromedin B (NMB-R).
- Neuromedin K (NK-3R).
- Neuropeptide Y types 1 to 6.
- Neurotensin (NT-R).
- 30 - Octopamine (tyramine), from insects.
- Odorants [11].
- Opioids delta-, kappa- and mu-types [12].
- Oxytocin (OT-R).

- Platelet activating factor (PAF-R).
- Prostacyclin.
- Prostaglandin D2.
- Prostaglandin E2, EP1 to EP4 subtypes.
- 5 - Prostaglandin F2.
- Purinoreceptors (ATP) [13].
- Somatostatin types 1 to 5.
- Substance-K (NK-2R).
- Substance-P (NK-1R).
- 10 - Thrombin.
- Thromboxane A2.
- Thyrotropin (TSH-R) [10].
- Thyrotropin releasing factor (TRH-R).
- Vasopressin V1a, V1b and V2.
- 15 - Visual pigments (opsins and rhodopsin) [14].
- Proto-oncogene mas.
- A number of orphan receptors (whose ligand is not known) from mammals and birds.
- Caenorhabditis elegans putative receptors C06G4.5, C38C10.1, C43C3.2, T27D1.3 and ZC84.4.
- 20 - Three putative receptors encoded in the genome of cytomegalovirus: US27, US28, and UL33.
- ECRF3, a putative receptor encoded in the genome of herpesvirus saimiri.

25 The structure of all these receptors is thought to be identical. They have seven hydrophobic regions, each of which most probably spans the membrane. The N-terminus is located on the extracellular side of the membrane and is often glycosylated, while the C-terminus is cytoplasmic and generally phosphorylated. Three extracellular loops alternate with three intracellular

30 loops to link the seven transmembrane regions. Most, but not all of these receptors, lack a signal peptide. The most conserved parts of these proteins are the transmembrane regions and the first two cytoplasmic loops. A conserved acidic-Arg-aromatic triplet is present in the N-terminal extremity of the

second cytoplasmic loop [15] and could be implicated in the interaction with G proteins.

To detect this widespread family of proteins, a pattern that contains the conserved triplet and that also spans the major part of the third transmembrane helix has been developed.

-Consensus pattern: [GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-
[LIVMNQGA]-x(2)-
[LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)-
[LIVM]

[1] Strosberg A.D.

Eur. J. Biochem. 196:1-10(1991).

[2] Kerlavage A.R.

Curr. Opin. Struct. Biol. 1:394-401(1991).

[3] Probst W.C., Snyder L.A., Schuster D.I., Brosius J., Sealfon S.C.

DNA Cell Biol. 11:1-20(1992).

[4] Savarese T.M., Fraser C.M.

Biochem. J. 283:1-9(1992).

[5] Branchek T.

Curr. Biol. 3:315-317(1993).

[6] Stiles G.L.

J. Biol. Chem. 267:6451-6454(1992).

[7] Friell T., Kobilka B.K., Lefkowitz R.J., Caron M.G.

Trends Neurosci. 11:321-324(1988).

[8] Stevens C.F.

Curr. Biol. 1:20-22(1991).

[9] Sakurai T., Yanagisawa M., Masaki T.

Trends Pharmacol. Sci. 13:103-107(1992).

[10] Salesse R., Remy J.J., Levin J.M., Jallal B., Garnier J.

Biochimie 73:109-120(1991).

[11] Lancet D., Ben-Arie N.

Curr. Biol. 3:668-674(1993).

[12] Uhl G.R., Childers S., Pasternak G.

Trends Neurosci. 17:89-93(1994).

[13] Barnard E.A., Burnstock G., Webb T.E.

5 Trends Pharmacol. Sci. 15:67-70(1994).

[14] Applebury M.L., Hargrave P.A.

Vision Res. 26:1881-1895(1986).

[15] Attwood T.K., Eliopoulos E.E., Findlay J.B.C.

Gene 98:153-159(1991).

10

(7tm 1) Visual pigments (opsins) retinal binding site

Visual pigments [1,2] are the light-absorbing molecules that mediate vision.

They consist of an apoprotein, opsin, covalently linked to the chromophore

cis-retinal. Vision is effected through the absorption of a photon by cis-

retinal which is isomerized to trans-retinal. This isomerization leads to a

change of conformation of the protein. Opsins are integral membrane proteins

with seven transmembrane regions that belong to family 1 of G-protein coupled receptors.

20

In vertebrates four different pigments are generally found. Rod cells, which mediate vision in dim light, contain the pigment rhodopsin. Cone cells, which function in bright light, are responsible for color vision and contain three or more color pigments (for example, in mammals: red, blue and green).

25

In *Drosophila*, the eye is composed of 800 facets or ommatidia. Each ommatidium contains eight photoreceptor cells (R1-R8): the R1 to R6 cells are outer cells, R7 and R8 inner cells. Each of the three types of cells (R1-R6, R7 and R8) expresses a specific opsin.

30

Proteins evolutionary related to opsins include squid retinochrome, also known as retinal photoisomerase, which converts various isomers of retinal into 11-cis retinal and mammalian retinal pigment epithelium (RPE) RGR [3], a protein that may also act in retinal isomerization.

The attachment site for retinal in the above proteins is a conserved lysine residue in the middle of the seventh transmembrane helix. The pattern that had been developed includes this residue.

5

-Consensus pattern: [LIVMWAC]-[PGC]-x(3)-[SAC]-K-[STALIMR]-[GSACPNV]-
[STACP]-
x(2)-[DENF]-[AP]-x(2)-[IY]
[K is the retinal binding site]

10

[1] Applebury M.L., Hargrave P.A.
Vision Res. 26:1881-1895(1986).

[2] Fryxell K.J., Meyerowitz E.M.
J. Mol. Evol. 33:367-378(1991).

15

[3] Shen D., Jiang M., Hao W., Tao L., Salazar M., Fong H.K.W.
Biochemistry 33:13117-13125(1994).

The following descriptions of protein family functions are not provided by the Pfam or Prosite databases.

20

740. BAH

BAH domain. Number of members: 65

25

[1] Medline: 97074677. Molecular cloning of polybromo, a nuclear protein containing multiple domains including five bromodomains, a truncated HMG-box, and two repeats of a novel domain. Nicolas RH, Goodwin GH; Gene 1996;175:233-240.

[2] Medline: 99198739. The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation. Callebaut I, Courvalin J-C,

30

Mornon JP; FEBS letts 1999;446:189-193.

741. ELM2.

ELM2 domain. The ELM2 (Egl-27 and MTA1 homology 2) domain is a small domain of unknown function. Number of members: 10

- 5 742. Euk proin. EUKARYOTIC_PORIN The major protein of the outer mitochondrial membrane of eukaryotes is a porin that forms a voltage-dependent anion-selective channel (VDAC) that behaves as a general diffusion pore for small hydrophilic molecules [1 to 4]. The channel adopts an open conformation at low or zero membrane potential and a closed conformation at potentials above 30-40 mV.

- 10 This protein contains about 280 amino acids and its sequence is composed of between 12 to 16 beta-strands that span the mitochondrial outer membrane. Yeast contains two members of this family (genes POR1 and POR2); vertebrates have at least three members (genes VDAC1, VDAC2 and VDAC3) [5].

A conserved region located at the C-terminal part of these proteins was selected as a signature pattern.

Consensus pattern[YH]-x(2)-D-[SPCAD]-x-[STA]-x(3)-[TAG]-[KR]-[LIVMF]-[DNSTA]-[DNS]-x(4)-[GSTAN]-[LIVMA]-x-[LIVMY]

- [1] Benz R. Biochim. Biophys. Acta 1197:167-196(1994).
 [2] Manella C.A. Trends Biochem. Sci. 17:315-320(1992).
 [3] Dihanich M. Experientia 46:146-153(1990).
 [4] Forte M., Guy H.R., Mannella C.A. J. Bioenerg. Biomembr. 19:341-350(1987).
 [5] Sampson M.J., Lovell R.S., Davison D.B., Craigen W.J. Genomics 36:192-196(1996).

743. Glyco hydor 19

Chitinases family 19 signatures

cross-reference(s) CHITINASE_19_1, CHITINASE_19_2

- 30 Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1]. Chitinases of family 19 (also known as classes IA or I and IB or II)

603

are enzymes from plants that function in the defense against fungal and insect pathogens by destroying their chitin-containing cell wall. Class IA/I and IB/II enzymes differ in the presence (IA/I) or absence (IB/II) of a N-terminal chitin-binding domain (see the relevant entry <PDOC00025>). The catalytic domain of these enzymes consist of about 220 to 230 amino acid residues.

Two highly conserved regions were selected as signature patterns, the first one is located in the N-terminal section and contains one of the six cysteines which are conserved in most, if not all, of these chitinases and which is probably involved in a disulfide bond.

Consensus pattern C-x(4,5)-F-Y-[ST]-x(3)-[FY]-[LIVMF]-x-A-x(3)-[YF]-x(2)-F-[GSA]
Consensus pattern [LIVM]-[GSA]-F-x-[STAG](2)-[LIVMFY]-W-[FY]-W-[LIVM]

[1]Flach J., Pilet P.-E., Jolles P. *Experientia* 48:701-716(1992).

[2] Henrissat B. *Biochem. J.* 280:309-316(1991).

744. MBD

Methyl-CpG binding domain

The Methyl-CpG binding domain (MBD) binds to DNA that contains one or more symmetrically methylated CpGs [1]. DNA methylation in animals is associated with alterations in chromatin structure and silencing of gene expression. MBD has negligible non-specific affinity for DNA. In vitro foot-printing with MeCP2 showed the MBD can protect a 12 nucleotide region surrounding a methyl CpG pair [1]. MBDs are found in several Methyl-CpG binding proteins and also DNA demethylase [2]. Number of members: 11

[1]Medline: 94232813. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. Nan X, Meehan RR, Bird A; *Nucleic Acids Res* 1993;21:4886-4892.

[2]Medline: 99158138. A mammalian protein with specific demethylase activity for mCpG DNA. Bhattacharya SK, Ramchandani S, Cervoni N, Szyf M; *Nature* 1999;397:579-583.

745. Peptidase C1

Eukaryotic thiol (cysteine) proteases active sites

cross-reference(s) THIOI_PROTEASE_CYS; THIOI_PROTEASE_HIS;
THIOI_PROTEASE_ASN

Eukaryotic thiol proteases (EC 3.4.22.-) [1] are a family of proteolytic enzymes which contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. The proteases which are currently known to belong to this family are listed below (references are only provided for recently determined sequences).

- Vertebrate lysosomal cathepsins B (EC 3.4.22.1), H (EC 3.4.22.16), L (EC 3.4.22.15), and S (EC 3.4.22.27) [2].

- Vertebrate lysosomal dipeptidyl peptidase I (EC 3.4.14.1) (also known as cathepsin C) [2].

- Vertebrate calpains (EC 3.4.22.17). Calpains are intracellular calcium- activated thiol protease that contain both a N-terminal catalytic domain and a C-terminal calcium-binding domain.

- Mammalian cathepsin K, which seems involved in osteoclastic bone resorption [3].

- Human cathepsin O [4].

- Bleomycin hydrolase. An enzyme that catalyzes the inactivation of the antitumor drug BLM (a glycopeptide).

- Plant enzymes: barley aleurain (EC 3.4.22.16), EP-B1/B4; kidney bean EP-C1, rice bean SH-EP; kiwi fruit actinidin (EC 3.4.22.14); papaya latex papain (EC 3.4.22.2), chymopapain (EC 3.4.22.6), caricain (EC 3.4.22.30), and proteinase IV (EC 3.4.22.25); pea turgor-responsive protein 15A; pineapple stem bromelain (EC 3.4.22.32); rape COT44; rice oryzain alpha, beta, and gamma; tomato low-temperature induced, Arabidopsis thaliana A494, RD19A and RD21A.

- House-dust mites allergens DerP1 and EurM1.

- Cathepsin B-like proteinases from the worms *Caenorhabditis elegans* (genes gcp-1, cpr-3, cpr-4, cpr-5 and cpr-6), *Schistosoma mansoni* (antigen SM31) and *Japonica* (antigen SJ31), *Haemonchus contortus* (genes AC-1 and AC-2), and *Ostertagia ostertagi* (CP-1 and CP-3).

- Slime mold cysteine proteinases CP1 and CP2.

- Cruzipain from *Trypanosoma cruzi* and *brucei*.

- Throphozoite cysteine proteinase (TCP) from various *Plasmodium* species.

- Proteases from *Leishmania mexicana*, *Theileria annulata* and *Theileria parva*.

605

- Baculoviruses cathepsin-like enzyme (v-cath).
- Drosophila small optic lobes protein (gene sol), a neuronal protein that contains a calpain-like domain.
- Yeast thiol protease BLH1/YCP1/LAP3.
- 5 - Caenorhabditis elegans hypothetical protein C06G4.2, a calpain-like protein.

Two bacterial peptidases are also part of this family:

- Aminopeptidase C from Lactococcus lactis (gene pepC) [5].
- 10 - Thiol protease tpr from Porphyromonas gingivalis.

Three other proteins are structurally related to this family, but may have lost their proteolytic activity.

5 - Soybean oil body protein P34. This protein has its active site cysteine replaced by a glycine.

- Rat testin, a sertoli cell secretory protein highly similar to cathepsin L but with the active site cysteine is replaced by a serine. Rat testin should not be confused with mouse testin which is a LIM-domain protein (see <PDOC00382>).

20 - Plasmodium falciparum serine-repeat protein (SERA), the major blood stage antigen. This protein of 111 Kd possesses a C-terminal thiol-protease-like domain [6], but the active site cysteine is replaced by a serine.

The sequences around the three active site residues are well conserved and can be used as signature patterns.

25 Consensus pattern Q-x(3)-[GE]-x-C-[YW]-x(2)-[STAGC]-[STAGCV] [C is the active site residue]

Note the residue in position 4 of the pattern is almost always cysteine; the only exceptions are calpains (Leu), bleomycin hydrolase (Ser) and yeast YCP1 (Ser). Note the residue in position 30 5 of the pattern is always Gly except in papaya protease IV where it is Glu.

Consensus pattern [LIVMGSTAN]-x-H-[GSACE]-[LIVM]-x-[LIVMAT](2)-G-x-[GSADNH] [H is the active site residue]

606

Consensus pattern[FYCH]-[WI]-[LIVT]-x-[KRQAG]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-[LFYW]-[LIVMFYG]-x-[LIVMF] [N is the active site residue]

Note these proteins belong to family C1 (papain-type) and C2 (calpains) in the classification of peptidases [7,E1].

5

[1]Dufour E. Biochimie 70:1335-1342(1988).

[2]Kirschke H., Barrett A.J., Rawlings N.D. Protein Prof. 2:1587-1643(1995).

[3]Shi G.-P., Chapman H.A., Bhairi S.M., Deleeuw C., Reddy V.Y., Weiss S.J. FEBS Lett. 357:129-134(1995).

10

[4]Velasco G., Ferrando A.A., Puente X.S., Sanchez L.M., Lopez-Otin C. J. Biol. Chem. 269:27136-27142(1994).

[5]Chapot-Chartier M.P., Nardi M., Chopin M.C., Chopin A., Gripon J.C. Appl. Environ. Microbiol. 59:330-333(1993).

[6]Higgins D.G., McConnell D.J., Sharp P.M. Nature 340:604-604(1989).

5

[7]Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

746. Peptidase M22

Glycoprotease family signature cross-reference(s) GLYCOPROTEASE

20

Glycoprotease (GCP) (EC 3.4.24.57) [1], or o-sialoglycoprotein endopeptidase, is a metalloprotease secreted by *Pasteurella haemolytica* which specifically cleaves O-sialoglycoproteins such as glycophorin A. The sequence of GCP is highly similar to the following uncharacterized proteins:

25

- *Escherichia coli* hypothetical protein ygiD (ORF-X).

- *Bacillus subtilis* hypothetical protein ydiE.

- *Mycobacterium leprae* hypothetical protein U229E.

- *Mycobacterium tuberculosis* hypothetical protein MtCY78.10.

- *Synechocystis* strain PCC 6803 hypothetical protein slr0807.

30

- *Methanococcus jannaschii* hypothetical protein MJ1130.

- *Haloarcula marismortui* hypothetical protein in HSH 3'region.

- Yeast hypothetical protein YKR038c.

- Yeast hypothetical protein QRI7.

One of the conserved regions contains two conserved histidines. It is possible that this region is involved in coordinating a metal ion such as zinc.

- 5 Consensus pattern[KR]-[GSAT]-x(4)-[FYWLH]-[DQNGK]-x-P-x-[LIVMFY]-x(3)-H-x(2)-[AG]-H-[LIVM]

Note these proteins belong to family M22 in the classification of peptidases [2,E1].

- 10 [1]Abdullah K.M., Lo R.Y.C., Mellors A. J. Bacteriol. 173:5597-5603(1991).
[2]Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

747. SAM. SAM domain (Sterile alpha motif)

15 It has been suggested that SAM is an evolutionarily conserved protein binding domain that is involved in the regulation of numerous developmental processes in diverse eukaryotes. The SAM domain can potentially function as a protein interaction module through its ability to homo- and heterooligomerise with other SAM domains. Number of members: 81

20 [1]Medline: 96100659 SAM: A novel motif in yeast sterile alpha and Drosophila polyhomeotic proteins Ponting CP; Prot Sci 1995;4:1928-1930.

[2]Medline: 97160498 SAM as a protein interaction domain involved in developmental regulation. Shultz J, Ponting CP, Hofmann K, Bork P; Prot Sci 1997;6:249-253.

[3]Medline: 99101382 The crystal structure of an Eph receptor SAM domain reveals a

- 25 mechanism for modular dimerization. Reference Author: Stapleton D, Balan I, Pawson T, Sicheri F; Nat Struct Biol 1999;6:44-49.

748. Tyrosinase signatures cross-reference(s) TYROSINASE_1; TYROSINASE_2

- 30 Tyrosinase (EC 1.14.18.1) [1] is a copper monooxygenases that catalyzes the hydroxylation of monophenols and the oxidation of o-diphenols to o-quinols. This enzyme, found in prokaryotes as well as in eukaryotes, is involved in the formation of pigments such as melanins and other polyphenolic compounds.

Tyrosinase binds two copper ions (CuA and CuB). Each of the two copper ion has been shown [2] to be bound by three conserved histidines residues. The regions around these copper-binding ligands are well conserved and also shared by some hemocyanins, which are copper-containing oxygen carriers from the hemolymph of many molluscs and arthropods [3,4].

At least two proteins related to tyrosinase are known to exist in mammals:

- TRP-1 (TYRP1) [5], which is responsible for the conversion of 5,6-dihydroxyindole-2-carboxylic acid (DHICA) to indole-5,6-quinone-2-carboxylic acid.
- TRP-2 (TYRP2) [6], which is the melanogenic enzyme DOPAchrome tautomerase (EC 5.3.3.12) that catalyzes the conversion of DOPAchrome to DHICA. TRP-2 differs from tyrosinases and TRP-1 in that it binds two zinc ions instead of copper [7].

Other proteins that belong to this family are:

- Plants polyphenol oxidases (PPO) (EC 1.10.3.1) which catalyze the oxidation of mono- and o-diphenols to o-diquinones [8].
- *Caenorhabditis elegans* hypothetical protein C02C2.1.

Two signature patterns for tyrosinase and related proteins have been derived. The first one contains two of the histidines that bind CuA, and is located in the N-terminal section of tyrosinase. The second pattern contains a histidine that binds CuB, that pattern is located in the central section of the enzyme.

Consensus pattern H-x(4,5)-F-[LIVMFTP]-x-[FW]-H-R-x(2)-[LM]-x(3)-E
[The two H's are copper ligands]

Consensus pattern D-P-x-F-[LIVMFYW]-x(2)-H-x(3)-D [H is a copper ligand]

[1] Lerch K. Prog. Clin. Biol. Res. 256:85-98(1988).

- [2]Jackman M.P., Hajnal A., Lerch K. Biochem. J. 274:707-713(1991).
- [3]Linzen B. Naturwissenschaften 76:206-211(1989).
- [4]Lang W.H., van Holde K.E. Proc. Natl. Acad. Sci. U.S.A. 88:244-248(1991).
- [5]Kobayashi T., Urabe K., Winder A., Jimenez-Cervantes C., Imokawa G., Brewington T., Solano F., Garcia-Borrón J.C., Hearing V.J. EMBO J. 13:5818-5825(1994).
- [6]Jackson I.J., Chambers D.M., Tsukamoto K., Copeland N.G., Gilbert D.J., Jenkins N.A., Hearing V. EMBO J. 11:527-535(1992).
- [7]Solano F., Martínez-Liarte J.H., Jimenez-Cervantes C., Garcia-Borrón J.C., Lozano J.A. Biochem. Biophys. Res. Commun. 204:1243-1250(1994).
- [8]Cary J.W., Lax A.R., Flurkey W.H. Plant Mol. Biol. 20:245-253(1992).

749. (Mur Ligase) Folylpolyglutamate synthase signatures

Folylpolyglutamate synthase (EC 6.3.2.17) (FPGS) [1] is the enzyme of folate metabolism that catalyzes ATP-dependent addition of glutamate moieties to tetrahydrofolate.

Its sequence is moderately conserved between prokaryotes (gene folC) and eukaryotes. We developed two signature patterns based on the conserved regions which are rich in glycine residues and could play a role in the catalytical activity and/or in substrate binding.

Description of pattern(s) and/or profile(s)

Consensus pattern[LIVMFY]-x-[LIVM]-[STAG]-G-T-[NK]-G-K-x-[ST]-x(7)-[LIVM](2)-x(3)-[GSK]

Consensus pattern[LIVMFY](2)-E-x-G-[LIVM]-[GA]-G-x(2)-D-x-[GST]-x-[LIVM](2)

- [1]Shane B., Garrow T., Brenner A., Chen L., Choi Y.J., Hsu J.C., Stover P. Adv. Exp. Med. Biol. 338:629-634(1993).

750. (Peptidase M3) Neutral zinc metallopeptidases, zinc-binding region signature

The majority of zinc-dependent metallopeptidases (with the notable exception of the carboxypeptidases) share a common pattern of primary structure [1,2,3] in the part of their

sequence involved in the binding of zinc, and can be grouped together as a superfamily, known as the metzincins, on the basis of this sequence similarity. They can be classified into a number of distinct families [4,E1] which are listed below along with the proteases which are currently known to belong to these families.

5

Family M1

- Bacterial aminopeptidase N (EC 3.4.11.2) (gene pepN).
- Mammalian aminopeptidase N (EC 3.4.11.2).
- Mammalian glutamyl aminopeptidase (EC 3.4.11.7) (aminopeptidase A). It may play a role in regulating growth and differentiation of early B-lineage cells.
- Yeast aminopeptidase yscII (gene APE2).
- Yeast alanine/arginine aminopeptidase (gene AAP1).
- Yeast hypothetical protein YIL137c.
- Leukotriene A-4 hydrolase (EC 3.3.2.6). This enzyme is responsible for the hydrolysis of an epoxide moiety of LTA-4 to form LTB-4; it has been shown that it binds zinc and is capable of peptidase activity.

10

15

Family M2

- Angiotensin-converting enzyme (EC 3.4.15.1) (dipeptidyl carboxypeptidase I) (ACE) the enzyme responsible for hydrolyzing angiotensin I to angiotensin II. There are two forms of ACE: a testis-specific isozyme and a somatic isozyme which has two active centers.

20

Family M3

- Thimet oligopeptidase (EC 3.4.24.15), a mammalian enzyme involved in the cytoplasmic degradation of small peptides.
- Neurolysin (EC 3.4.24.16) (also known as mitochondrial oligopeptidase M or microsomal endopeptidase).
- Mitochondrial intermediate peptidase precursor (EC 3.4.24.59) (MIP). It is involved the second stage of processing of some proteins imported in the mitochondrion.
- Yeast saccharolysin (EC 3.4.24.37) (proteinase yscD).
- Escherichia coli and related bacteria dipeptidyl carboxypeptidase (EC 3.4.15.5) (gene dcp).
- Escherichia coli and related bacteria oligopeptidase A (EC 3.4.24.70) (gene opdA or prIC).

25

30

- Yeast hypothetical protein YKL134c.

Family M4

- Thermostable thermolysins (EC 3.4.24.27), and related thermolabile neutral proteases (bacillolysins) (EC 3.4.24.28) from various species of *Bacillus*.

- Pseudolysin (EC 3.4.24.26) from *Pseudomonas aeruginosa* (gene *lasB*).

- Extracellular elastase from *Staphylococcus epidermidis*.

- Extracellular protease *prt1* from *Erwinia carotovora*.

- Extracellular minor protease *smp* from *Serratia marcescens*.

- Vibriolysin (EC 3.4.24.25) from various species of *Vibrio*.

- Protease *prtA* from *Listeria monocytogenes*.

- Extracellular proteinase *proA* from *Legionella pneumophila*.

Family M5

- Mycolysin (EC 3.4.24.31) from *Streptomyces cacaoi*.

Family M6

- Immune inhibitor A from *Bacillus thuringiensis* (gene *ina*). *Ina* degrades two classes of insect antibacterial proteins, attacins and cecropins.

Family M7

- *Streptomyces* extracellular small neutral proteases

Family M8

- Leishmanolysin (EC 3.4.24.36) (surface glycoprotein gp63), a cell surface protease from various species of *Leishmania*.

Family M9

- Microbial collagenase (EC 3.4.24.3) from *Clostridium perfringens* and *Vibrio alginolyticus*.

Family M10A

- Serralysin (EC 3.4.24.40), an extracellular metalloprotease from *Serratia*.

- Alkaline metalloproteinase from *Pseudomonas aeruginosa* (gene aprA).
- Secreted proteases A, B, C and G from *Erwinia chrysanthemi*.
- Yeast hypothetical protein YIL108w.

5 Family M10B

- Mammalian extracellular matrix metalloproteinases (known as matrixins) [5]: MMP-1 (EC 3.4.24.7) (interstitial collagenase), MMP-2 (EC 3.4.24.24) (72 Kd gelatinase), MMP-9 (EC 3.4.24.35) (92 Kd gelatinase), MMP-7 (EC 3.4.24.23) (matrylsin), MMP-8 (EC 3.4.24.34) (neutrophil collagenase), MMP-3 (EC 3.4.24.17) (stromelysin-1), MMP-10 (EC 3.4.24.22) (stromelysin-2), and MMP-11 (stromelysin-3), MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).
- Sea urchin hatching enzyme (envelysin) (EC 3.4.24.12). A protease that allows the embryo to digest the protective envelope derived from the egg extracellular matrix.
- Soybean metalloendoproteinase 1.

5 Family M11

- *Chlamydomonas reinhardtii* gamete lytic enzyme (GLE).

20 Family M12A

- Astacin (EC 3.4.24.21), a crayfish endoprotease.
- Meprin A (EC 3.4.24.18), a mammalian kidney and intestinal brush border metalloendopeptidase.
- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity. The *Drosophila* homolog of BMP-1 is the dorsal-ventral patterning protein tolloid.
- Blastula protease 10 (BP10) from *Paracentrotus lividus* and the related protein SpAN from *Strongylocentrotus purpuratus*.
- *Caenorhabditis elegans* protein toh-2.
- *Caenorhabditis elegans* hypothetical protein F42A10.8.
- Choriolysins L and H (EC 3.4.24.67) (also known as embryonic hatching proteins LCE and HCE) from the fish *Oryzias latipes*. These proteases participate in the breakdown of the egg envelope, which is derived from the egg extracellular matrix, at the time of hatching.

Family M12B

- Snake venom metalloproteinases [6]. This subfamily mostly groups proteases that act in hemorrhage. Examples are: adamalysin II (EC 3.4.24.46), atrolysin C/D (EC 3.4.24.42), atrolysin E (EC 3.4.24.44), fibrolase (EC 3.4.24.72), trimereylisin I (EC 3.4.25.52) and II (EC 3.4.25.53).
- Mouse cell surface antigen MS2.

Family M13

- Mammalian neprilysin (EC 3.4.24.11) (neutral endopeptidase) (NEP).
- Endothelin-converting enzyme 1 (EC 3.4.24.71) (ECE-1), which process the precursor of endothelin to release the active peptide.
- Kell blood group glycoprotein, a major antigenic protein of erythrocytes. The Kell protein is very probably a zinc endopeptidase.
- Peptidase O from *Lactococcus lactis* (gene pepO).

Family M27

- Clostridial neurotoxins, including tetanus toxin (TeTx) and the various botulinum toxins (BoNT). These toxins are zinc proteases that block neurotransmitter release by proteolytic cleavage of synaptic proteins such as synaptobrevins, syntaxin and SNAP-25 [7,8].

Family M30

- *Staphylococcus hyicus* neutral metalloprotease.

Family M32

- Thermostable carboxypeptidase 1 (EC 3.4.17.19) (carboxypeptidase Taq), an enzyme from *Thermus aquaticus* which is most active at high temperature.

Family M34

- Lethal factor (LF) from *Bacillus anthracis*, one of the three proteins composing the anthrax toxin.

Family M35

- Deuterolysin (EC 3.4.24.39) from *Penicillium citrinum* and related proteases from various species of *Aspergillus*.

5 Family M36

- Extracellular elastinolytic metalloproteinases from *Aspergillus*.

From the tertiary structure of thermolysin, the position of the residues acting as zinc ligands and those involved in the catalytic activity are known. Two of the zinc ligands are histidines which are very close together in the sequence; C-terminal to the first histidine is a glutamic acid residue which acts as a nucleophile and promotes the attack of a water molecule on the carbonyl carbon of the substrate. A signature pattern which includes the two histidine and the glutamic acid residues is sufficient to detect this superfamily of proteins.

Description of pattern(s) and/or profile(s)

Consensus pattern[GSTALIVN]-x(2)-H-E-[LIVMFYW]-{DEHRKP}-H-x-[LIVMFYWGSPQ] [The two H's are zinc ligands] [E is the active site residue]

Sequences known to belong to this class detected by the patternALL, except for members of families M5, M7 and M11.

Other sequence(s) detected in SWISS-PROT55; including *Neurospora crassa* conidiation-specific protein 13 which could be a zinc-protease.

[1]Jongeneel C.V., Bouvier J., Bairoch A.
FEBS Lett. 242:211-214(1989).

[2]Murphy G.J.P., Murphy G., Reynolds J.J.
FEBS Lett. 289:4-7(1991).

[3]Bode W., Grams F., Reinemer P., Gomis-Rueth F.-X., Baumann U., McKay D.B., Stoecker W.
Zoology 99:237-246(1996).

[4]Rawlings N.D., Barrett A.J.
Meth. Enzymol. 248:183-228(1995).

[5]Woessner J. Jr.

FASEB J. 5:2145-2154(1991).

[6]Hite L.A., Fox J.W., Bjarnason J.B.

[7]Montecucco C., Schiavo G.

5 Trends Biochem. Sci. 18:324-327(1993).

[8]Niemann H., Blasi J., Jahn R.

Trends Cell Biol. 4:179-185(1994).

10 751. PseudoU_synt_1

tRNA pseudouridine synthase is involved in the formation of pseudouridine at the anticodon stem and loop of transfer-RNAs Pseudouridine is an isomer of uridine (5-(beta-D-ribofuranosyl) uracil, and is the most abundant modified nucleoside found in all cellular RNAs. The TruA-like proteins also exhibit a conserved sequence with a strictly conserved aspartic acid, likely involved in catalysis. Number of members: 25

[1]Medline: 98254513. Transfer RNA-pseudouridine synthetase Pus1 of *Saccharomyces cerevisiae* contains one atom of zinc essential for its native conformation and tRNA recognition. Arluison V, Hountondji C, Robert B, Grosjean H; *Biochemistry* 1998;37:7268-7276.

752. EPSP synthase signatures

EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase) (EC 2.5.1.19) catalyzes the sixth step in the biosynthesis from chorismate of the aromatic amino acids (the shikimate pathway) in bacteria (gene *aroA*), plants and fungi (where it is part of a multifunctional enzyme which catalyzes five consecutive steps in this pathway) [1]. EPSP synthase has been extensively studied as it is the target of the potent herbicide glyphosate which inhibits the enzyme.

The sequence of EPSP from various biological sources shows that the structure of the enzyme has been well conserved throughout evolution. Two conserved regions were selected as signature patterns. The first pattern corresponds to a region that is part of the active site and

616

which is also important for the resistance to glyphosate [2]. The second pattern is located in the C-terminal part of the protein and contains a conserved lysine which seems to be important for the activity of the enzyme.

5 Description of pattern(s) and/or profile(s)

Consensus pattern[LIVM]-x(2)-[GN]-N-[SA]-G-T-[STA]-x-R-x-[LIVMY]-x-[GSTA]

Consensus pattern[KR]-x-[KH]-E-[CST]-[DNE]-R-[LIVM]-x-[STA]-[LIVMC]-x(2)-[EN]-[LIVMF]-x-[KRA]-[LIVMF]-G

10

[1]Stallings W.C., Abdel-Megid S.S., Lim L.W., Shieh H.-S., Dayringer H.E., Leimgruber N.K., Stegeman R.A., Anderson K.S., Sikorski J.A., Padgett S.R., Kishore G.M. Proc. Natl. Acad. Sci. U.S.A. 88:5046-5050(1991).

[2]Padgett S.R., Re D.B., Gaser C.S., Eicholtz D.A., Frazier R.B., Hironaka C.M., Levine E.B., Shah D.M., Fraley R.T., Kishore G.M. J. Biol. Chem. 266:22364-22369(1991).

753. Glyco_hydro_18

Glycosyl hydrolases family 18. Number of members: 173

[1]Medline: 95219379. Crystal structure of a bacterial chitinase at 2.3 A resolution. Perrakis A, Tews I, Dauter Z, Oppenheim AB, Chet I, Wilson KS, Vorgias CE; Structure 1994;2:1169-1180.

25 754. Esterase

Putative esterase

This family contains Esterase D Swiss:P10768. However it is not clear if all members of the family have the same function. This family is possibly related to the COesterase family.

Number of members: 36

30

755. (HMA) Heavy-metal-associated domain

A conserved domain of about 30 amino acid residues has been found [1] in a number of proteins that transport or detoxify heavy metals. This domain contains two conserved cysteines that could be involved in the binding of these metals. The domain has been termed Heavy-Metal-Associated (HMA). It has been found in:

- A variety of cation transport ATPases (E1-E2 ATPases) (see <PDOC00139>). The human copper ATPases ATP7A and ATP7B which are respectively involved in Menke's and Wilson's diseases. ATP7A and ATP7B both contain 6 tandem copies of the HMA domain. The copper ATPases CCC2 from budding yeast, copA from *Enterococcus faecalis* and synA from *Synechococcus* contain one copy of the HMA domain. The cadmium ATPases cadA from *Bacillus firmus* and from plasmid pI258 from *Staphylococcus aureus* also contain a single HMA domain, while a chromosomal *Staphylococcus aureus* cadA contains two copies. Other, less characterized ATPases that contain the HMA domain are: fixI from *Rhizobium meliloti*, pacS from *Synechococcus* strain PCC 7942), *Mycobacterium leprae* ctpA and ctpB and *Escherichia coli* hypothetical protein yhhO. In all these ATPases the HMA domain(s) are located in the N-terminal section.
- Mercuric reductase (EC 1.16.1.1) (gene merA) which is generally encoded by plasmids carried by mercury-resistant Gram-negative bacteria. Mercuric reductase is a class-1 pyridine nucleotide-disulphide oxidoreductase (see <PDOC00073>). There is generally one HMA domain (with the exception of a chromosomal merA from *Bacillus* strain RC607 which has two) in the N-terminal part of merA.
- Mercuric transport protein periplasmic component (gene merP), also encoded by plasmids carried by mercury-resistant Gram-negative bacteria. It seems to be a mercury scavenger that specifically binds to one Hg(2+) ion and which passes it to the mercuric reductase via the merT protein. The N-terminal half of merP is a HMA domain.
- *Helicobacter pylori* copper-binding protein copP.
- Yeast protein ATX1 [2], which could act in the transport and/or partitioning of copper.

The consensus pattern for HMA spans the complete domain.

Description of pattern(s) and/or profile(s)

Consensus pattern[LIVN]-x(2)-[LIVMFA]-x-C-x-[STAGCDNH]-C-x(3)-[LIVFG]-x(3)-
[LIV]-x(9,11)-[IVA]-x-[LVFYS] [The two C's probably bind metals]

[1]Bull P.C., Cox D.W. Trends Genet. 10:246-252(1994).

5 [2]Lin S.-J., Culotta V.L. Proc. Natl. Acad. Sci. U.S.A. 92:3784-3788(1995).

756. (Peptidase M10) Matrixins cysteine switch

PROSITE cross-reference(s): CYSTEINE_SWITCH

Mammalian extracellular matrix metalloproteinases (EC 3.4.24.-), also known as matrixins

10 [1] (see <PDOC00129>), are zinc-dependent enzymes. They are secreted by cells in an
inactive form (zymogen) that differs from the mature enzyme by the presence of an N-
terminal propeptide. A highly conserved octapeptide is found two residues downstream of
the C-terminal end of the propeptide. This region has been shown to be involved in
autoinhibition of matrixins [2,3]; a cysteine within the octapeptide chelates the active site
5 zinc ion, thus inhibiting the enzyme. This region has been called the 'cysteine switch' or
'autoinhibitor region'.

A cysteine switch has been found in the following zinc proteases:

- MMP-1 (EC 3.4.24.7) (interstitial collagenase).
- 20 - MMP-2 (EC 3.4.24.24) (72 Kd gelatinase).
- MMP-3 (EC 3.4.24.17) (stromelysin-1).
- MMP-7 (EC 3.4.24.23) (matrilysin).
- MMP-8 (EC 3.4.24.34) (neutrophil collagenase).
- MMP-9 (EC 3.4.24.35) (92 Kd gelatinase).
- 25 - MMP-10 (EC 3.4.24.22) (stromelysin-2).
- MMP-11 (EC 3.4.24.-) (stromelysin-3).
- MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).
- MMP-13 (EC 3.4.24.-) (collagenase 3).
- MMP-14 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 1).
- 30 - MMP-15 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 2).
- MMP-16 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 3).
- Sea urchin hatching enzyme (EC 3.4.24.12) (envelysin) [4].
- Chlamydomonas reinhardtii gamete lytic enzyme (GLE) [5].

Description of pattern(s) and/or profile(s)

Consensus pattern P-R-C-[GN]-x-P-[DR]-[LIVSAPKQ] [C chelates the zinc ion]

- 5 [1]Woessner J. Jr. FASEB J. 5:2145-2154(1991).
 [2]Sanchez-Lopez R., Nicholson R., Gesnel M.C., Matrisian L.M., Breathnach R. J. Biol. Chem. 263:11892-11899(1988).
 [3]Park A.J., Matrisian L.M., Kells A.F., Pearson R., Yuan Z., Navre M. J. Biol. Chem. 266:1584-1590(1991).
 10 [4]Lepage T., Gache C. EMBO J. 9:3003-3012(1990).
 [5]Kinoshita T., Fukuzawa H., Shimada T., Saito T., Matsuda Y. Proc. Natl. Acad. Sci. U.S.A. 89:4693-4697(1992).

757. (Peptidase S8) Serine proteases, subtilase family, active sites

PROSITE cross-reference(s): PS00136; SUBTILASE_ASP, PS00137; SUBTILASE_HIS, PS00138; SUBTILASE_SER

Subtilases [1,2] are an extensive family of serine proteases whose catalytic activity is provided by a charge relay system similar to that of the trypsin family of serine proteases but which evolved by independent convergent evolution. The sequence around the residues involved in the catalytic triad (aspartic acid, serine and histidine) are completely different from that of the analogous residues in the trypsin serine proteases and can be used as signatures specific to that category of proteases.

The subtilase family currently includes the following proteases:

- 25 - Subtilisins (EC 3.4.21.62), these alkaline proteases from various *Bacillus* species have been the target of numerous studies in the past thirty years.
 - Alkaline elastase YaB from *Bacillus* sp. (gene ale).
 - Alkaline serine exoprotease A from *Vibrio alginolyticus* (gene proA).
 - Aqualysin I from *Thermus aquaticus* (gene pstI).
 30 - AspA from *Aeromonas salmonicida*.
 - Bacillopeptidase F (esterase) from *Bacillus subtilis* (gene bpf).
 - C5A peptidase from *Streptococcus pyogenes* (gene scpA).
 - Cell envelope-located proteases PI, PII, and PIII from *Lactococcus lactis*.

- Extracellular serine protease from *Serratia marcescens*.
- Extracellular protease from *Xanthomonas campestris*.
- Intracellular serine protease (ISP) from various *Bacillus*.
- Minor extracellular serine protease epr from *Bacillus subtilis* (gene epr).
- 5 - Minor extracellular serine protease vpr from *Bacillus subtilis* (gene vpr).
- Nisin leader peptide processing protease nisP from *Lactococcus lactis*.
- Serotype-specific antigene 1 from *Pasteurella haemolytica* (gene ssa1).
- Thermitase (EC 3.4.21.66) from *Thermoactinomyces vulgaris*.
- Calcium-dependent protease from *Anabaena variabilis* (gene prcA).
- 10 - Halolysin from halophilic bacteria sp. 172p1 (gene hly).
- Alkaline extracellular protease (AEP) from *Yarrowia lipolytica* (gene xpr2).
- Alkaline proteinase from *Cephalosporium acremonium* (gene alp).
- Cerevisin (EC 3.4.21.48) (vacuolar protease B) from yeast (gene PRB1).
- Cuticle-degrading protease (pr1) from *Metarhizium anisopliae*.
- 15 - KEX-1 protease from *Kluyveromyces lactis*.
- Kexin (EC 3.4.21.61) from yeast (gene KEX-2).
- Oryzin (EC 3.4.21.63) (alkaline proteinase) from *Aspergillus* (gene alp).
- Proteinase K (EC 3.4.21.64) from *Tritirachium album* (gene proK).
- Proteinase R from *Tritirachium album* (gene proR).
- 20 - Proteinase T from *Tritirachium album* (gene proT).
- Subtilisin-like protease III from yeast (gene YSP3).
- Thermomycolin (EC 3.4.21.65) from *Malbranchea sulfurea*.
- Furin (EC 3.4.21.85), neuroendocrine convertases 1 to 3 (NEC-1 to -3) and PACE4 protease from mammals, other vertebrates, and invertebrates. These proteases are involved
- 25 in the processing of hormone precursors at sites comprised of pairs of basic amino acid residues [3].
- Tripeptidyl-peptidase II (EC 3.4.14.10) (tripeptidyl aminopeptidase) from Human.
- Prestalk-specific proteins tagB and tagC from slime mold [4]. Both proteins consist of two domains: a N-terminal subtilase catalytic domain and a C-terminal ABC transporter domain
- 30 (see <PDOC00185>).

Description of pattern(s) and/or profile(s)

621

Consensus pattern[STAlV]-x-[LIVMF]-[LIVM]-D-[DSTA]-G-[LIVMFC]-x(2,3)-[DNH] [D is the active site residue]

Consensus patternH-G-[STM]-x-[VIC]-[STAGC]-[GS]-x-[LIVMA]-[STAGCLV]-[SAGM] [H is the active site residue]

5 Consensus patternG-T-S-x-[SA]-x-P-x(2)-[STAVC]-[AG] [S is the active site residue]

Note if a protein includes at least two of the three active site signatures, the probability of it being a serine protease from the subtilase family is 100%

Note these proteins belong to family S8 in the classification of peptidases [5,E1].

10

[1]Siezen R.J., de Vos W.M., Leunissen J.A.M., Dijkstra B.W. Protein Eng. 4:719-737(1991).

[2]Siezen R.J. (In) Proceeding subtilisin symposium, Hamburg, (1992).

[3]Barr P.J. Cell 66:1-3(1991).

[4]Shaulsky G., Kuspa A., Loomis W.F.; Genes Dev. 9:1111-1122(1995).

[5]Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

758. (SSB) Single-strand binding protein family signatures

PROSITE cross-reference(s): PS00735; SSB_1,PS00736; SSB_2

The Escherichia coli single-strand binding protein [1] (gene ssb), also known as the helix-destabilizing protein, is a protein of 177 amino acids. It binds tightly, as a homotetramer, to single-stranded DNA (ss-DNA) and plays an important role in DNA replication, recombination and repair.

25

Closely related variants of SSB are encoded in the genome of a variety of large self-transmissible plasmids. SSB has also been characterized in bacteria such as Proteus mirabilis or Serratia marcescens.

30 Eukaryotic mitochondrial proteins that bind ss-DNA and are probably involved in mitochondrial DNA replication are structurally and evolutionary related to prokaryotic SSB. Proteins currently known to belong to this subfamily are listed below [2].

- Mammalian protein Mt-SSB (P16).

- Xenopus Mt-SSBs and Mt-SSBr.
- Drosophila MtSSB.
- Yeast protein RIM1.

- 5 Two signature patterns have been developed for these proteins. The first is a conserved region in the N-terminal section of the SSB's. The second is a centrally located region which, in Escherichia coli SSB, is known to be involved in the binding of DNA.

Description of pattern(s) and/or profile(s)

10 Consensus pattern[LIVMF]-[NST]-[KRT]-[LIVM]-x-[LIVMF](2)-G-[NHRK]-[LIVM]-[GST]-x-[DET]

Consensus patternT-x-W-[HY]-[RNS]-[LIVM]-x-[LIVMF]-[FY]-[NGKR]

[1]Meyer R.R., Laine P.S. Microbiol. Rev. 54:342-380(1990).

15 [2]Stroumbakis N.D., Li Z., Tolias P.P. Gene 143:171-177(1994).

759. KDPG and KHG aldolases active site signatures

PROSITE cross-reference(s): PS00159; ALDOLASE_KDPG_KHG_1, PS00160;
ALDOLASE_KDPG_KHG_2

20 4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (KHG-aldolase) catalyzes the interconversion of 4-hydroxy-2-oxoglutarate into pyruvate and glyoxylate. Phospho-2-dehydro-3-deoxygluconate aldolase (EC 4.1.2.14) (KDPG-aldolase) catalyzes the interconversion of 6-phospho-2-dehydro-3-deoxy-D-gluconate into pyruvate and
25 glyceraldehyde 3-phosphate.

These two enzymes are structurally and functionally related [1]. They are both homotrimeric proteins of approximately 220 amino-acid residues. They are class I aldolases whose catalytic mechanism involves the formation of a Schiff-base intermediate between the substrate and
30 the epsilon-amino group of a lysine residue. In both enzymes, an arginine is required for catalytic activity.

Two signature patterns were developed for these enzymes. The first one contains the active site arginine and the second, the lysine involved in the Schiff-base formation.

Description of pattern(s) and/or profile(s)

- 5 Consensus pattern G-[LIVM]-x(3)-E-[LIV]-T-[LF]-R [R is the active site residue]
 Consensus pattern G-x(3)-[LIVMF]-K-[LF]-F-P-[SA]-x(3)-G [K is involved in Schiff-base formation]

[1] Vlahos C J., Dekker E.E. J. Biol. Chem. 263:11683-11691(1988).

10

760. AP endonucleases family 1 signatures. PROSITE cross-reference(s): PS00726;
 AP_NUCLEASE_F1_1, PS00727; AP_NUCLEASE_F1_2, PS00728;
 AP_NUCLEASE_F1_3

15

DNA damaging agents such as the antitumor drugs bleomycin and neocarzinostatin or those that generate oxygen radicals produce a variety of lesions in DNA. Amongst these is base-loss which forms apurinic/apyrimidinic (AP) sites or strand breaks with atypical 3'termini. DNA repair at the AP sites is initiated by specific endonuclease cleavage of the phosphodiester backbone. Such endonucleases are also generally capable of removing blocking groups from the 3'terminus of DNA strand breaks.

20

AP endonucleases can be classified into two families on the basis of sequence similarity. Family 1 groups the enzymes listed below [1].

- 25 - Escherichia coli exonuclease III (EC 3.1.11.2) (gene xthA).
 - Streptococcus pneumoniae and Bacillus subtilis exonuclease A (gene exoA).
 - Mammalian AP endonuclease 1 (AP1) (EC 4.2.99.18).
 - Drosophila recombination repair protein 1 (gene Rrp1).
 - Arabidopsis thaliana apurinic endonuclease-redox protein (gene arp).

30

Except for Rrp1 and arp, these enzymes are proteins of about 300 amino-acid residues. Rrp1 and arp both contain additional and unrelated sequences in their N-terminal section (about 400 residues for Rrp1 and 270 for arp).

Three signature patterns were developed for this family of enzymes. The patterns are based on the most conserved regions. The first pattern contains a glutamate which has been shown [2], in the Escherichia coli enzyme to bind a divalent metal ion such as magnesium or manganese

Consensus pattern[APF]-D-[LIVMF](2)-x-[LIVM]-Q-E-x-K [E binds a divalent metal ion]
Consensus patternD-[ST]-[FY]-R-[KH]-x(7,8)-[FYW]-[ST]-[FYW](2)
Consensus patternN-x-G-x-R-[LIVM]-D-[LIVMFYH]-x-[LV]-x-S

[1] Barzilay G., Hickson I.S. BioEssays 17:713-719(1995).

[2] Mol C.D., Kuo C.-F., Thayer M.M., Cunningham R.P., Tainer J.A. Nature 374:381-386(1995).

761. (ER)Enhancer of rudimentary signature, PROSITE cross-reference(s): PS01290; ER

The Drosophila protein 'enhancer of rudimentary' (gene (e(r)) is a small protein of 104 residues whose function is not yet clear. From an evolutionary point of view, it is highly conserved [1] and has been found to exist in probably all multicellular eukaryotic organisms. It has been proposed that this protein plays a role in the cell cycle.

A conserved region in the central part of the protein was selected as as signaure pattern.

Consensus patternY-D-I-[SA]-x-L-[FY]-x-F-[IV]-D-x(3)-D-[LIV]-S

[1] Gelsthorpe M., Pulumati M., McCallum C., Dang-Vu K., Tsubota S.I. Gene 186:189-195(1997).

762. (ETF alpha) Electron transfer flavoprotein alpha-subunit signature, PROSITE cross-reference(s): PS00696; ETF_ALPHA

The electron transfer flavoprotein (ETF) [1,2] serves as a specific electron acceptor for various mitochondrial dehydrogenases. ETF transfers electrons to the main respiratory

625

chain via ETF-ubiquinone oxidoreductase. ETF is an heterodimer that consist of an alpha and a beta subunit and which bind one molecule of FAD per dimer. A similar system also exists in some bacteria.

- 5 The alpha subunit of ETF is a protein of about 32 Kd which is structurally related to the bacterial nitrogen fixation protein fixB which could play a role in a redox process and feed electrons to ferredoxin.

Other related proteins are:

10

- Escherichia coli hypothetical protein ydiR.
- Escherichia coli hypothetical protein ygcQ.

A highly conserved region which is located in the C-terminal section was selected as a signature pattern for these proteins.

15

Consensus pattern [LI]-Y-[LIVM]-[AT]-x-G-[IV]-[SD]-G-x-[IV]-Q-H-x(2)-G-x(6)-[IV]-x-A-[IV]-N

20

- [1] Finocchiaro G., Ikeda Y., Ito M., Tanaka K. Prog. Clin. Biol. Res. 321:637-652(1990).
- [2] Tsai M.H., Saier M.H. Jr. Res. Microbiol. 146:397-404(1995).

763. (lectin c) C-type lectin domain signature and profile

PROSITE cross-reference(s): PS00615; C_TYPE_LLECTIN_1, PS50041;

25

C_TYPE_LLECTIN_2

A number of different families of proteins share a conserved domain which was first characterized in some animal lectins and which seem to function as a calcium-dependent carbohydrate-recognition domain [1,2,3]. This domain, which is known as the C-type lectin domain (CTL) or as the carbohydrate-recognition domain (CRD), consists of about 110 to 130 residues. There are four cysteines which are perfectly conserved and involved in two disulfide bonds. A schematic representation of the CTL domain is shown below.

30

```

+-----+
|   |
xcxxxxxxxxxxxxCxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxCxxxxWxCxxxxC
|   |   | *****|*
+-----+   +-----+

```

'C': conserved cysteine involved in a disulfide bond.

'c': optional cysteine involved in a disulfide bond.

'*': position of the pattern.

The categories of proteins, in which the CTL domain has been found, are listed below.

Type-II membrane proteins where the CTL domain is located at the C-terminal extremity of the proteins:

- Asialoglycoprotein receptors (ASGPR) (also known as hepatic lectins) [4]. The ASGPR's mediate the endocytosis of plasma glycoproteins to which the terminal sialic acid residue in their carbohydrate moieties has been removed.

- Low affinity immunoglobulin epsilon Fc receptor (lymphocyte IgE receptor), which plays an essential role in the regulation of IgE production and in the differentiation of B cells.

- Kupffer cell receptor. A receptor with an affinity for galactose and fucose, that could be involved in endocytosis.

- A number of proteins expressed on the surface of natural killer T-cells: NKG2, NKR-P1, YE1/88 (Ly-49), CD69 and on B-cells: CD72, LyB-2. The CTL-domain in these proteins is distantly related to other CTL-domains; it is unclear whether they are likely to bind carbohydrates.

Proteins that consist of an N-terminal collagenous domain followed by a CTL- domain [5], these proteins are sometimes called 'collectins':

- Pulmonary surfactant-associated protein A (SP-A). SP-A is a calcium-dependent protein that binds to surfactant phospholipids and contributes to lower the surface tension at the air-liquid interface in the alveoli of the

mammalian lung.

- Pulmonary surfactant-associated protein D (SP-D).
- Conglutinin, a calcium-dependent lectin-like protein which binds to a yeast cell wall extract and to immune complexes through the complement component (iC3b).
- Mannan-binding proteins (MBP) (also known as mannose-binding proteins). MBP's bind mannose and N-acetyl-D-glucosamine in a calcium-dependent manner.
- Bovine collectin-43 (CL-43).

Selectins (or LEC-CAM) [6,7]. Selectins are cell adhesion molecules implicated in the interaction of leukocytes with platelets or vascular endothelium. Structurally, selectins consist of a long extracellular domain, followed by a transmembrane region and a short cytoplasmic domain. The extracellular domain is itself composed of a CTL-domain, followed by an EGF-like domain and a variable number of SCR/Sushi repeats. Known selectins are:

- Lymph node homing receptor (also known as L-selectin, leukocyte adhesion molecule-1, (LAM-1), leu-8, gp90-mel, or LECAM-1)
- Endothelial leukocyte adhesion molecule 1 (ELAM-1, E-selectin or LECAM-2). The ligand recognized by ELAM-1 is sialyl-Lewis x.
- Granule membrane protein 140 (GMP-140, P-selectin, PADGEM, CD62, or LECAM-3). The ligand recognized by GMP-140 is Lewis x.

Large proteoglycans that contain a CTL-domain followed by one copy of a SCR/ Sushi repeat, in their C-terminal section:

- Aggrecan (cartilage-specific proteoglycan core protein). This proteoglycan is a major component of the extracellular matrix of cartilaginous tissues where it has a role in the resistance to compression.
- Brevican.
- Neurocan.
- Versican (large fibroblast proteoglycan), a large chondroitin sulfate

proteoglycan that may play a role in intercellular signalling.

In addition to the CTL and Sushi domains, these proteins also contain, in their N-terminal domain, an Ig-like V-type region, two or four link domains (see <PDOC00955>) and up to
5 two EGF-like repeats.

Two type-I membrane proteins:

- 10 - Mannose receptor from macrophages. This protein mediates the endocytosis of glycoproteins by macrophages in several recognition and uptake processes. Its extracellular section consists of a fibronectin type II domain followed by eight tandem repeats of the CTL domain.
- 180 Kd secretory phospholipase A2 receptor (PLA2-R). A protein whose structure is highly similar to that of the mannose receptor.
- 15 - DEC-205 receptor. This protein is used by dendritic cells and thymic epithelial cells to capture and endocytose diverse carbohydrate-binding antigens and direct them to antigen-processing cellular compartments. DEC-205 extracellular section consists of a fibronectin type II domain followed by ten tandem repeats of the CTL domain.
- 20 - Silk moth hemocytin, an humoral lectin which is involved in a self-defence mechanism. It is composed of 2 FA58C domains (see <PDOC00988>), a CTL domain, 2 VWFC domains (see <PDOC00928>), and a CTCK (see <PDOC00912>).

Various other proteins that uniquely consist of a CTL domain:

- 25 - Invertebrate soluble galactose-binding lectins. A category to which belong a humoral lectin from a flesh fly; echinoidin, a lectin from the coelomic fluid of a sea urchin; BRA-2 and BRA-3, two lectins from the coelomic fluid of a barnacle, a lectin from the tunicate Polyandrocarpa misakiensis and a
30 newt oviduct lectin. The physiological importance of these lectins is not yet known but they may play an important role in defense mechanisms.
- Pancreatic stone protein (PSP) (also known as pancreatic thread protein (PTP), or reg), a protein that might act as an inhibitor of spontaneous

calcium carbonate precipitation.

- Pancreatitis associated protein (PAP), a protein that might be involved in the control of bacterial proliferation.
- Tetranectin, a plasma protein that binds to plasminogen and to isolated kringle 4.
- Eosinophil granule major basic protein (MBP), a cytotoxic protein.
- A galactose specific lectin from a rattlesnake.
- Two subunits of a coagulation factor IX/factor X-binding protein (IX/X-bp), a snake venom anticoagulant protein which binds with factors IX and X in the presence of calcium.
- Two subunits of a phospholipase A2 inhibitor from the plasma of a snake (PLI-A and PLI-B).
- A lipopolysaccharide-binding protein (LPS-BP) from the hemolymph of a cockroach [8].
- Sea raven antifreeze protein (AFP) [9].

As a signature pattern for this domain, the C-terminal region with its three conserved cysteines was selected.

Consensus pattern C-[LIVMFYATG]-x(5,12)-[WL]-x-[DNSR]-x(2)-C-x(5,6)-[FYWLIVSTA]-[LIVMSTA]-C [The three C's are involved in disulfide bonds]

Note all CTL domains have five Trp residues before the second Cys, with the exception of tunicate lectin and cockroach LPS-BP which have Leu.

Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.

[1] Drickamer K. J. Biol. Chem. 263:9557-9560(1988).

[2] Drickamer K. Prog. Nucleic Acid Res. Mol. Biol. 45:207-232(1993).

- [3] Drickamer K. Curr. Opin. Struct. Biol. 3:393-400(1993).
- [4] Spiess M. Biochemistry 29:10009-10018(1990).
- [5] Weis W.I., Kahn R., Fourme R., Drickamer K., Hendrickson W.A. Science 254:1608-1615(1991).
- 5 [6] Siegelman M. Curr. Biol. 1:125-128(1991).
- [7] Lasky L.A. Science 238:964-969(1992).
- [8] Jomori T., Natori S. J. Biol. Chem. 266:13318-13323(1991).
- [9] Ng N.F.L., Hew C.-L. J. Biol. Chem. 267:16069-16075(1992).

10 764. (SRCR) Speract receptor repeated domain signature
 PROSITE cross-reference(s): PS00420; SPERACT_RECEPTOR,

The receptor for the sea urchin egg peptide speract is a transmembrane glycoprotein of 500 amino acid residues [1]. Structurally it consists of a large extracellular domain of 450 residues, followed by a transmembrane region and a small cytoplasmic domain of 12 amino acids. The extracellular domain contains four repeats of a 115 amino acids domain. There are 17 positions that are perfectly conserved in the four repeats, among them are six cysteines, six glycines, and three glutamates.

Such a domain is also found, once, in the C-terminal section of mammalian macrophage scavenger receptor type I [2], a membrane glycoproteins implicated in the pathologic deposition of cholesterol in arterial walls during atherogenesis.

The signature pattern that was derived spans part of the N-terminal section of the domain and contains 8 of the 17 conserved residues.

Consensus pattern G-x(5)-G-x(2)-E-x(6)-W-G-x(2)-C-x(3)-[FYW]-x(8)-C-x(3)-G

[1] Dangott J.J., Jordan J.E., Bellet R.A., Garbers D.L. Proc. Natl. Acad. Sci. U.S.A. 86:2128-2132(1989).

[2] Freeman M., Ashkenas J., Rees D.J., Kingsley D.M., Copeland N.G., Jenkins N.A., Krieger M. Proc. Natl. Acad. Sci. U.S.A. 87:8810-8814(1990).

765. Bac_surface_Ag

Bacterial surface antigen

This entry includes the following surface antigens; D15 antigen from *H.influenzae*, OMA87 from *P.multocida*, OMP85 from *N.meningitidis* and *N.gonorrhoeae*. Number of members:

5 14

[1] Medline: 95255676. The sequencing of the 80-kDa D15 protective surface antigen of *Haemophilus influenzae*. Flack FS, Loosmore S, Chong P, Thomas WR; Gene 1995;156:97-99.

10 [2] Medline: 96333354. Cloning, sequencing, expression, and protective capacity of the oma87 gene encoding the *Pasteurella multocida* 87-kilodalton outer membrane antigen. Ruffolo CG, Adler B; Infect Immun 1996;64:3161-3167.

766. BRCA1 C Terminus (BRCT) domain

5 The BRCT domain is found predominantly in proteins involved in cell cycle checkpoint functions responsive to DNA damage. It has been suggested that the Retinoblastoma protein contains a divergent BRCT domain, this has not been included in this family. The BRCT domain of XRCC1 forms a homodimer in the crystal structure Medline:99016060. This suggests that pairs of BRCT domains

20 associate as homo- or heterodimers. Number of members: 131

[1] Medline: 96259550. BRCA1 protein products ...Functional motifs... Koonin EV, Altschul SF, Bork P; Nature Genet 1996;13:266-268.

25 [2] Medline: 97153217. From BRCA1 to RAP1: A widespread BRCT module closely associated with DNA repair Callebaut I, Mornon JP; Febs lett 1997;400:25-30.

[3] Medline: 97186552. A superfamily of conserved domains in DNA damage responsive cell cycle checkpoint proteins Bork P, Hofmann K, Bucher P, Neuwald AF, Altschul SF, Koonin EV; Faseb J 1997;11:68-76.

30 [4] Medline: 97402527. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ; Nucleic Acids Res 1997;25:3389-3402.

[5] Medline: 99016060. Structure of an XRCC1 BRCT domain: a new protein-protein interaction module. Zhang X, Morera S, Bates PA, Whitehead PC, Coffe AI, Hainbucher K, Nash RA, Sternberg MJ, Lindahl T, Freemont PS;

5 767. Kappa casein

Kappa-casein is a mammalian milk protein involved in a number of important physiological processes. In the gut, the ingested protein is split into an insoluble peptide (para kappa-casein) and a soluble hydrophilic glycopeptide (caseinomacropeptide). Caseinomacropeptide is responsible for increased efficiency of digestion, prevention of neonate hypersensitivity to ingested proteins, and inhibition of gastric pathogens. Number of members: 56

[1] Medline: 98072500. Nucleotide sequence evolution at the kappa-casein locus: evidence for positive selection within the family Bovidae. Ward TJ, Honeycutt RL, Derr JN; Genetics 1997;147:1863-1872.

768. Chitinases family 18 active site

PROSITE cross-reference(s) CHITINASE_18

Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1]. Chitinases of family 18 (also known as classes III or V) groups a variety of proteins:

a) Chitinases from:

- Prokaryotes such as Alteromonas, Bacillus, Serratia, Streptomyces, etc.
- Plants such as Arabidopsis, cucumber, bean, tobacco, etc.
- Fungi such as Aphanocladium, Rhizopus, Saccharomyces, etc.
- Nematode (Brugia malayi).
- Insects (Manduca sexta).
- Baculoviruses (Autographa Californica Nuclear Polyhedrosis virus).

b) Other proteins:

- Hevamine, a rubber tree protein with chitinase and lysozyme activities.
- Kluyveromyces lactis killer toxin alpha subunit, which acts as a chitinase.
- Flavobacterium and Streptomyces endo-beta-N-acetylglucosaminidases (EC 3.2.1.96).
- Mammalian di-N-acetylchitobiase which is involved in the degradation of asparagine-linked glycoproteins.
- Human cartilage glycoprotein Gp-39.
- Jack bean concanavalin B (conB), a protein that has lost its catalytic activity.

Site directed mutagenesis experiments [3] and crystallographic data [4,5] have shown that a conserved glutamate is involved in the catalytic mechanism and probably acts as a proton donor. This glutamate is at the extremity of the best conserved region in these proteins.

Consensus pattern[LIVMFY]-[DN]-G-[LIVMF]-[DN]-[LIVMF]-[DN]-x-E [E is the active site residue]

- [1] Flach J., Pilet P.-E., Jolles P. Experientia 48:701-716(1992).
- [2] Henrissat B. Biochem. J. 280:309-316(1991).
- [3] Watanabe T., Kohori K., Miyashita K., Fujii T., Sakai H., Uchida M., Tanaka H. J. Biol. Chem. 268:18567-18572(1993).
- [4] Perrakis A., Tews I., Dauter Z., Oppenheim A.B., Chet I., Wilson K.S., Vorgias C.E. Structure 2:1169-1180(1994).
- [5] van Scheltinga A.C.T., Kalk K.H., Beintema J.J., Dijkstra B.W. Structure 2:1181-1189(1994).

769. gag_p17. gag gene protein p17 (matrix protein).

The matrix protein forms an icosahedral shell associated with the inner membrane of the mature immunodeficiency virus. Number of members: 1598

[1] Medline: 95055757. Three-dimensional structure of the human immunodeficiency virus type 1 matrix protein. Massiah MA, Starich MR, Paschall C, Summers MF, Christensen AM, Sundquist WI; J Mol Biol 1994;244:198-223.

770. GDA1/CD39 family of nucleoside phosphatases signature

634

PROSITE cross-reference(s); GDA1_CD39_NTPASE

A number of nucleoside diphosphate and triphosphate hydrolases as well as some yet uncharacterized proteins have been found to belong to the same family [1, 2]. This family currently consist of:

- 5 - Yeast guanosine-diphosphatase (EC 3.6.1.42) (GDPase) (gene GDA1). GDA1 is a golgi integral membrane enzyme that catalyzes the hydrolysis of GDP to GMP.
- Potato apyrase (EC 3.6.1.5) (adenosine diphosphatase) (ADPase). Apyrase acts on both ATP and ADP to produce AMP.
- Mammalian vascular ATP-diphosphohydrolase (EC 3.6.1.5) (also known as lymphoid cell activation antigen CD39).
- 10 - Toxoplasma gondii nucleoside-triphosphatases (EC 3.6.1.15) (NTPase). NTPase hydrolyses various nucleoside triphosphates to produce the corresponding nucleoside mono- and diphosphates. This enzyme is secreted into the invaded host cell into the parasitophorous vacuole, a specialized compartment where the parasite intracellularly resides.
- Pea nucleoside-triphosphatases (EC 3.6.1.15) (NTPase).
- Caenorhabditis elegans hypothetical protein C33H5.14.
- Caenorhabditis elegans hypothetical protein R07E4.4.
- 20 - Yeast chromosome V hypothetical protein YER005w.

The above uncharacterized proteins all seem to be membrane-bound.

All these proteins share a number of conserved domains. The best conserved of these domains have been selected. It is located in the central section of the proteins.

Consensus pattern[LIVM]-x-G-x(2)-E-G-x-[FY]-x-[FW]-[LIVA]-[TAG]-x-N-[HY]

- 30 [1] Handa M., Guidotti G. Biochem. Biophys. Res. Commun. 218:916-923(1996).
- [2] Vasconcelos E.G., Ferreira S.T., de Carvalho T.M.U., de Souza W., Kettlun A.M., Mancilla M., Valenzuela M.A., Verjovski-Almeida S. J. Biol. Chem. 271:22139-22145(1996).

771. GTP cyclohydrolase I signatures

PROSITE cross-reference(s); GTP_CYCLOHYDROL_1_1, GTP_CYCLOHYDROL_1_2

GTP cyclohydrolase I (EC 3.5.4.16) catalyzes the biosynthesis of formic acid and
 5 dihydroneopterin triphosphate from GTP. This reaction is the first step in the biosynthesis of
 tetrahydrofolate in prokaryotes, of tetrahydrobiopterin in vertebrates, and of pteridine-
 containing pigments in insects.

10 GTP cyclohydrolase I is a protein of from 190 to 250 amino acid residues. The comparison
 of the sequence of the enzyme from bacterial and eukaryotic sources shows that the
 structure of this enzyme has been extremely well conserved throughout evolution [1].

Two conserved regions were selected as signature patterns. The first contains a perfectly
 conserved tetrapeptide which is part of the GTP-binding pocket [2], the second region also
 5 contains conserved residues involved in GTP-binding.

Consensus pattern[DEN]-[LIVM](2)-x(2)-[KRNQ]-[DEN]-[LIVM]-x(3)-[ST]-x-C-E- H-H
 Consensus pattern[SA]-x-[RK]-x-Q-[LIVM]-Q-E-[RN]-[LI]-[TSN]

20 [1] Maier J., Witter K., Guetlich M., Ziegler I., Werner T., Ninnemann H. Biochem.
 Biophys. Res. Commun. 212:705-711(1995).

[2] Nar H., Huber R., Meining W., Schmid C., Weinkauff S., Bacher A. Structure 3:459-
 466(1995).

25 772. IlvC. Acetohydroxy acid isomeroreductase

Acetohydroxy acid isomeroreductase catalyses the conversion of acetohydroxy acids into
 dihydroxy valerates. This reaction is the second in the synthetic pathway of the essential
 branched side chain amino acids valine and isoleucine. Number of members: 29

30 [1] Medline: 97361822. The crystal structure of plant acetohydroxy acid isomeroreductase
 complexed with NADPH, two magnesium ions and a herbicidal transition state analog
 determined at 1.65 Å resolution. Biou V, Dumas R, Cohen-Addad C, Douce R, Job D, Pebay-
 Peyroula E; EMBO J 1997;16:3405-3415.

773. Prokaryotic membrane lipoprotein lipid attachment site

PROSITE cross-reference(s); PROKAR_LIPOPROTEIN

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

- Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp).
- Escherichia coli lipoprotein-28 (gene nlpA).
- Escherichia coli lipoprotein-34 (gene nlpB).
- Escherichia coli lipoprotein nlpC.
- Escherichia coli lipoprotein nlpD.
- Escherichia coli osmotically inducible lipoprotein B (gene osmB).
- Escherichia coli osmotically inducible lipoprotein E (gene osmE).
- Escherichia coli peptidoglycan-associated lipoprotein (gene pal).
- Escherichia coli rare lipoproteins A and B (genes rplA and rplB).
- Escherichia coli copper homeostasis protein cutF (or nlpE).
- Escherichia coli plasmids traT proteins.
- Escherichia coli Col plasmids lysis proteins.
- A number of Bacillus beta-lactamases.
- Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA).
- Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB).
- Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7).
- Chlamydia trachomatis outer membrane protein 3 (gene omp3).
- Fibrobacter succinogenes endoglucanase cel-3.
- Haemophilus influenzae proteins Pal and Pcp.
- Klebsiella pullulunase (gene pulA).
- Klebsiella pullulunase secretion protein pulS.
- Mycoplasma hyorhina protein p37.
- Mycoplasma hyorhina variant surface antigens A, B, and C (genes vlpABC).
- Neisseria outer membrane protein H.8.
- Pseudomonas aeruginosa lipopeptide (gene lppL).

637

- *Pseudomonas solanacearum* endoglucanase egl.
- *Rhodopseudomonas viridis* reaction center cytochrome subunit (gene cytC).
- *Rickettsia* 17 Kd antigen.
- *Shigella flexneri* invasion plasmid proteins mxiJ and mxiM.
- 5 - *Streptococcus pneumoniae* oligopeptide transport protein A (gene amiA).
- *Treponema pallidum* 34 Kd antigen.
- *Treponema pallidum* membrane protein A (gene tmpA).
- *Vibrio harveyi* chitinase (gene chb).
- *Yersinia* virulence plasmid protein yscJ.

10

- Halocyanin from *Natrobacterium pharaonis* [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion).

From the precursor sequences of all these proteins, we derived a consensus pattern and a set of rules to identify this type of post-translational modification.

Consensus pattern{DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence.

[1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).

[2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).

[3] von Heijne G. Protein Eng. 2:531-534(1989).

25 [4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

774. Aminoacyl-transfer RNA synthetases class-II signatures

PROSITE cross-reference(s); AA_TRNA_LIGASE_II_1; AA_TRNA_LIGASE_II_2

30 Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are

generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure.

- 5 The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7].

10

Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns from two of these regions have been derived.

11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

5 Consensus pattern[FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern[GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]

[1]Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).

[2]Delarue M., Moras D. BioEssays 15:675-687(1993).

[3]Schimmel P. Trends Biochem. Sci. 16:1-3(1991).

[4]Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

[5]Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).

[6]Cusack S. Biochimie 75:1077-1081(1993).

25 [7]Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).

[8]Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

775. X. Trans-activation protein X

30 This protein is found in hepadnaviruses where it is indispensable for replication. Number of members: 91

776. Thymidylate synthase active site

Thymidylate synthase (EC 2.1.1.45) [1,2] catalyzes the reductive methylation of dUMP to dTMP with concomitant conversion of 5,10-methylenetetrahydrofolate to dihydrofolate. Thymidylate synthase plays an essential role in DNA synthesis and is an important target for certain chemotherapeutic drugs.

5 Thymidylate synthase is an enzyme of about 30 to 35 Kd in most species except in protozoan and plants where it exists as a bifunctional enzyme that includes a dihydrofolate reductase domain.

A cysteine residue is involved in the catalytic mechanism (it covalently binds the 5,6-dihydro-dUMP intermediate). The sequence around the active site of this enzyme is
10 conserved from phages to vertebrates.

Consensus pattern R-x(2)-[LIVM]-x(3)-[FW]-[QN]-x(8,9)-[LV]-x-P-C-[HAVM]-x(3)-[QMT]-[FYW]-x-[LV] [C is the active site residue]

15 [1] Benkovic S.J. Annu. Rev. Biochem. 49:227-251(1980).

[2] Ross P., O'Gara F., Condon S. Appl. Environ. Microbiol. 56:2156-2163(1990).

777. Glycosyl hydrolases family 31 signatures

It has been shown [1,2,3,E1] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

- Lysosomal alpha-glucosidase (EC 3.2.1.20) (acid maltase) is a vertebrate glycosidase active at low pH, which hydrolyzes alpha(1->4) and alpha(1->6) linkages in glycogen, maltose, and isomaltose.

- Alpha-glucosidase (EC 3.2.1.20) from the yeast *Candida tsukunbaensis*.

25 - Alpha-glucosidase (EC 3.2.1.20) (gene *malA*) from the archaebacteria *Sulfolobus solfataricus*.

- Intestinal sucrase-isomaltase (EC 3.2.1.48 / EC 3.2.1.10) is a vertebrate membrane-bound, multifunctional enzyme complex which hydrolyzes sucrose, maltose and isomaltose. The sucrase and isomaltase domains of the enzyme are homologous (41% of amino acid identity)
30 and have most probably evolved by duplication.

- Glucoamylase 1 (EC 3.2.1.3) (glucan 1,4-alpha-glucosidase) from various fungal species.

- Yeast hypothetical protein YBR229c.

- Fission yeast hypothetical protein SpAC30D11.01c.

An aspartic acid has been implicated [4] in the catalytic activity of sucrase, isomaltase, and lysosomal alpha-glucosidase. The region around this active residue is highly conserved and can be used as a signature pattern. A second region, which contains two conserved cysteines, has been used as an additional signature pattern.

5

Consensus pattern [GF]-[LIVMF]-W-x-D-M-[NSA]-E [D is the active site residue]

Consensus pattern G-[AV]-D-[LIVMTA]-C-G-[FY]-x(3)-[ST]-x(3)-L-C-x-R-W-x(2)-[LV]-[GSA]-[SA]-F-x-P-F-x-R-[DN]

10 [1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Kinsella B.T., Hogan S., Larkin A., Cantwell B.A. Eur. J. Biochem. 202:657-664(1991).

[3] Naim H.Y., Niermann T., Kleinhans U., Hollenberg C.P., Strasser A.W.M. FEBS Lett. 294:109-112(1991).

[4] Hermans M.M.P., Kroos M.A., van Beeumen J., Oostra B.A., Reuser A.J.J. J. Biol. Chem. 266:13507-13512(1991).

778. Urease signatures

Urease (EC 3.5.1.5) is a nickel-binding enzyme that catalyzes the hydrolysis of urea to carbon dioxide and ammonia [1]. Historically, it was the first enzyme to be crystallized (in 1926). It is mainly found in plant seeds, microorganisms and invertebrates. In plants, urease is a hexamer of identical chains. In bacteria [2], it consists of either two or three different subunits (alpha, beta and gamma).

Urease binds two nickel ions per subunit; four histidine, an aspartate and a carbamated-lysine serve as ligands to these metals; an additional histidine is involved in the catalytic mechanism [3].

As signatures for this enzyme, a region was selected that contains two histidine that bind one of the nickel ions and the region of the active site histidine.

Consensus pattern T-[AY]-[GA]-[GAT]-[LIVM]-D-x-H-[LIVM]-H-x(3)-P [The two H's bind nickel]

Consensus pattern [LIVM](2)-[CT]-H-[HN]-L-x(3)-[LIVM]-x(2)-D-[LIVM]-x-F-A [H is the active site residue]

[1] Takishima K., Suga T., Mamiya G. Eur. J. Biochem. 175:151-165(1988).

[2] Mobley H.L.T., Husinger R.P. Microbiol. Rev. 53:85-108(1989).

[3] Jabri E., Carr M.B., Hausinger R.P., Karplus P.A. Science 268:998-1004(1995).

5 779. Tyrosine specific protein phosphatases signature and profiles

Tyrosine specific protein phosphatases (EC 3.1.3.48) (PTPase) [1 to 5] are enzymes that catalyze the removal of a phosphate group attached to a tyrosine residue. These enzymes are very important in the control of cell growth, proliferation, differentiation and transformation. Multiple forms of PTPase have been characterized and can be classified into
10 two categories: soluble PTPases and transmembrane receptor proteins that contain PTPase domain(s). The currently known PTPases are listed below:

Soluble PTPases.

- PTPN1 (PTP-1B).

15 - PTPN2 (T-cell PTPase; TC-PTP).

- PTPN3 (H1) and PTPN4 (MEG), enzymes that contain an N-terminal band 4.1- like domain (see <PDOC00566>) and could act at junctions between the membrane and cytoskeleton.

- PTPN5 (STEP).

20 - PTPN6 (PTP-1C; HCP; SHP) and PTPN11 (PTP-2C; SH-PTP3; Syp), enzymes which contain two copies of the SH2 domain at its N-terminal extremity. The Drosophila protein corkscrew (gene csw) also belongs to this subgroup.

- PTPN7 (LC-PTP; Hematopoietic protein-tyrosine phosphatase; HePTP).

- PTPN8 (70Z-PEP).

25 - PTPN9 (MEG2).

- PTPN12 (PTP-G1; PTP-P19).

- Yeast PTP1.

- Yeast PTP2 which may be involved in the ubiquitin-mediated protein degradation pathway.

30 - Fission yeast pyp1 and pyp2 which play a role in inhibiting the onset of mitosis.

- Fission yeast pyp3 which contributes to the dephosphorylation of cdc2.

- Yeast CDC14 which may be involved in chromosome segregation.

- Yersinia virulence plasmid PTPases (gene yopH).

- Autographa californica nuclear polyhedrosis virus 19 Kd PTPase.

Dual specificity PTPases.

- DUSP1 (PTPN10; MAP kinase phosphatase-1; MKP-1); which dephosphorylates MAP kinase on both Thr-183 and Tyr-185.
- DUSP2 (PAC-1), a nuclear enzyme that dephosphorylates MAP kinases ERK1 and ERK2 on both Thr and Tyr residues.
- DUSP3 (VHR).
- DUSP4 (HVB2).
- DUSP5 (HVB3).
- DUSP6 (Pyst1; MKP-3).
- DUSP7 (Pyst2; MKP-X).
- Yeast MSG5, a PTPase that dephosphorylates MAP kinase FUS3.
- Yeast YVH1.
- Vaccinia virus H1 PTPase; a dual specificity phosphatase.

Receptor PTPases.

Structurally, all known receptor PTPases, are made up of a variable length extracellular domain, followed by a transmembrane region and a C-terminal catalytic cytoplasmic domain. Some of the receptor PTPases contain fibronectin type III (FN-III) repeats, immunoglobulin-like domains, MAM domains or carbonic anhydrase-like domains in their extracellular region. The cytoplasmic region generally contains two copies of the PTPase domain. The first seems to have enzymatic activity, while the second is inactive but seems to affect substrate specificity of the first. In these domains, the catalytic cysteine is generally conserved but some other, presumably important, residues are not.

In the following table, the domain structure of known receptor PTPases is shown:

Extracellular	Intracellular
-----	-----
Ig FN-3 CAH MAM	PTPase

Leukocyte common antigen (LCA) (CD45) 0 2 0 0 2

			643		
	Leukocyte antigen related (LAR)	3	8	0	0 2
	Drosophila DLAR	3	9	0	0 2
	Drosophila DPTP	2	2	0	0 2
	PTP-alpha (LRP)	0	0	0	0 2
5	PTP-beta	0	16	0	0 1
	PTP-gamma	0	1	1	0 2
	PTP-delta	0	>7	0	0 2
	PTP-epsilon	0	0	0	0 2
	PTP-kappa	1	4	0	1 2
10	PTP-mu	1	4	0	1 2
	PTP-zeta	0	1	1	0 2

PTPase domains consist of about 300 amino acids. There are two conserved cysteines, the second one has been shown to be absolutely required for activity. Furthermore, a number of conserved residues in its immediate vicinity have also been shown to be important.

A signature pattern was derived for PTPase domains centered on the active site cysteine.

There are three profiles for PTPases, the first one spans the complete domain and is not specific to any subtype. The second profile is specific to dual-specificity PTPases and the third one to the PTP subfamily.

Consensus pattern [LIVMF]-H-C-x(2)-G-x(3)-[STC]-[STAGP]-x-[LIVMFY] [C is the active site residue]

Notethe M-phase inducer phosphatases (cdc25-type phosphatase) are tyrosine- protein phosphatases that are not structurally related to the above PTPases.

Notethis documentation entry is linked to both a signature pattern and to profiles. As profiles are much more sensitive than the pattern, you should use them if you have access to the necessary software tools to do so.

[1] Fischer E.H., Charbonneau H., Tonks N.K. Science 253:401-406(1991).

[2] Charbonneau H., Tonks N.K. Annu. Rev. Cell Biol. 8:463-493(1992).

[3] Trowbridge I.S. J. Biol. Chem. 266:23517-23520(1991).

[4] Tonks N.K., Charbonneau H. Trends Biochem. Sci. 14:497-500(1989).

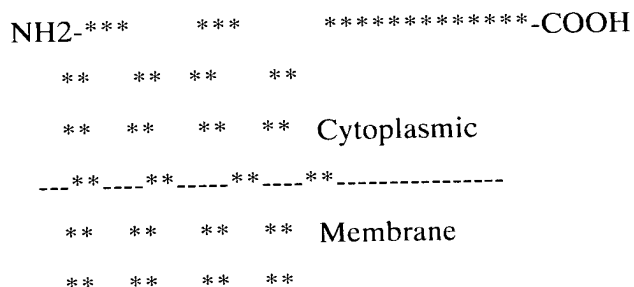
[5] Hunter T. Cell 58:1013-1016(1989).

780. Connexins signatures

Gap junctions [1] are specialized regions of the plasma membrane which consist of closely packed pairs of transmembrane channels, the connexons, through which small molecules diffuse from a cell to a neighboring cell. Each connexon is composed of an hexamer of an integral membrane protein which is often referred to as connexin. In a given species there are a number of different, yet structurally related, tissue specific, forms of connexins. The types of connexins which are currently known are listed below.

- Connexin 56 (Cx56).
- Connexin 50 (Cx50) (lens fiber protein MP70).
- Connexin 46 (Cx46) (alpha-3).
- Connexin 45 (Cx45) (alpha-6).
- Connexin 43 (Cx43) (alpha-1).
- Connexin 40 (Cx40) (alpha-5).
- Connexin 38 (Cx38) (alpha-2).
- Connexin 37 (Cx37) (alpha-4).
- Connexin 33 (Cx33) (alpha-7).
- Connexin 32 (Cx32) (beta-1).
- Connexin 31.1 (Cx31.1) (beta-4).
- Connexin 31 (Cx31) (beta-3).
- Connexin 30.3 (Cx30.3) (beta-5).
- Connexin 26 (Cx26) (beta-2).

Structurally the connexins consist of a short cytoplasmic N-terminal domain, followed by four transmembrane segments that delimit two extracellular and one cytoplasmic loops; the C-terminal domain is cytoplasmic and its length is variable (from 20 residues in Cx26 to 260 residues in Cx56). The schematic representation of this structure is shown below.



645

```

---**---**---**---**-----
**  **  **  ** Extracellular
**  **  **  **
**      **

```

5 The sequences of the two extracellular loops are well conserved. In both loops there are three conserved cysteines which are involved in disulfide bonds. A signature patterns from each of these two loop regions has been built.

10 Consensus pattern C-[DN]-T-x-Q-P-G-C-x(2)-V-C-[FY]-D [The three C's are involved in disulfide bonds] Consensus pattern C-x(3,4)-P-C-x(3)-[LIVM]-[DEN]-C-[FY]-[LIVM]-[SA]-[KR]-P [The three C's are involved in disulfide bonds]

[1] Goodenough D.A., Goliger J.A., Paul D.L. Annu. Rev. Biochem. 65:475-502(1996).

15 781. Gram-positive cocci surface proteins 'anchoring' hexapeptide

Surface proteins from Gram-positive cocci contains a conserved hexapeptide located a few residues downstream of a hydrophobic C-terminal membrane anchor region which is followed by a cluster of basic amino acids [1]. This structure is represented in the following schematic representation:

```

+-----+-----+-----+
| Variable length extracellular domain |H| Anchor |B|
+-----+-----+-----+

```

'H': conserved hexapeptide.

25 'B': cluster of basic residues.

It has been proposed that this hexapeptide sequence is responsible for a post-translational modification necessary for the proper anchoring of the proteins which bear it, to the cell wall. Proteins known to contain such hexapeptide are listed below:

- 30
- Aggregation substance from streptococcus faecalis (asa1).
 - C5a peptidase from Streptococcus pyogenes (scpA).
 - C protein alpha-antigen from Streptococcus agalactiae (bca).
 - Cell surface antigen I/II (PAC) from Streptococcus mutans.

- Dextranase from *Streptococcus downei* (dex).
- Fibronectin-binding protein from *Staphylococcus aureus* (fnbA).
- Fimbrial subunits from *Actinomyces naeslundii* and *viscosus*.
- IgA binding protein from *Streptococcus pyogenes* (arp4).
- 5 - IgA binding protein (B antigen) from *Streptococcus agalactiae* (bag).
- IgG binding proteins from *Streptococci* and *Staphylococcus aureus*.
- Internalin A from *Listeria monocytogenes* (inlA).
- M proteins from streptococci.
- Muramidase-released protein from *Streptococcus suis* (mrp).
- 10 - Nisin leader peptide processing protease from *Lactococcus lactis* (nisP).
- Protein A from *Staphylococcus aureus*.
- Trypsin-resistant surface T protein from streptococci.
- Wall-associated protein from *Streptococcus mutans* (wapA).
- Wall-associated serine proteinases from *Lactococcus lactis*.

Consensus pattern L-P-x-T-G-[STGAVDE]

[1] Schneewind O., Jones K.F., Fischetti V.A. J. Bacteriol. 172:3310-3317(1990).

782. Gamma-glutamyltranspeptidase signature

Gamma-glutamyltranspeptidase (EC 2.3.2.2) (GGT) [1] catalyzes the transfer of the gamma-glutamyl moiety of glutathione to an acceptor that may be an amino acid, a peptide or water (forming glutamate). GGT plays a key role in the gamma-glutamyl cycle, a pathway for the synthesis and degradation of glutathione. In prokaryotes and eukaryotes, it is an enzyme that consists of two polypeptide chains, a heavy and a light subunit, processed from a single chain precursor. The active site of GGT is known to be located in the light subunit.

The sequences of mammalian and bacterial GGT show a number of regions of high similarity [2]. *Pseudomonas cephalosporin acylases* (EC 3.5.1.-) that convert 7-beta-(4-carboxybutanamido)-cephalosporanic acid (GL-7ACA) into 7-aminocephalosporanic acid (7ACA) and glutaric acid are evolutionary related to GGT and also show some GGT activity [3]. Like GGT, these GL-7ACA acylases, are also composed of two subunits.

One of the conserved regions correspond to the N-terminal extremity of the mature light chains of these enzymes. This region has been used as a signature pattern.

Consensus pattern T-[STA]-H-x-[ST]-[LIVMA]-x(4)-G-[SN]-x-V-[STA]-x-T-x-T-[LIVM]-
[NE]-x(1,2)-[FY]-G

- 5 [1] Tate S.S., Meister A. Meth. Enzymol. 113:400-419(1985).
[2] Suzuki H., Kumagai H., Echigo T., Tochikura T. J. Bacteriol. 171:5169-5172(1989).
[3] Ishiye M., Niwa M. Biochim. Biophys. Acta 1132:233-239(1992).

783. Ferrochelatase signature

10 Ferrochelatase (EC 4.99.1.1) (protoheme ferro-lyase) [1,2] catalyzes the last step in heme biosynthesis: the chelation of a ferrous ion to proto-porphyrin IX, to form protoheme.

In eukaryotes, ferrochelatase is a mitochondrial protein bound to the inner membrane, whose active site faces the mitochondrial matrix. The mature form of eukaryotic ferrochelatase is composed of about 360 amino acids. In bacteria, ferrochelatase (gene hemH) [3] is a protein of from 310 to 380 amino acids.

The human autosomal dominant disease protoporphyria is due to the reduced activity of ferrochelatase.

The signature pattern for this enzyme is based on a conserved region which contains a histidine residue which could be involved in binding iron.

20 Consensus pattern [LIVMF](2)-x-[ST]-x-H-[GS]-[LIVM]-P-x(4,5)-[DENQKR]-x-G-[DP]-x(1,2)-Y

- [1] Labbe-Bois R. J. Biol. Chem. 265:7278-7283(1990).
25 [2] Brenner D.A., Frasier F. Proc. Natl. Acad. Sci. U.S.A. 88:849-853(1991).
[3] Miyamoto K., Nakahigashi K., Nishimura K., Inokuchi H. J. Mol. Biol. 219:393-398(1991).

784. Cellulose-binding domain, bacterial type

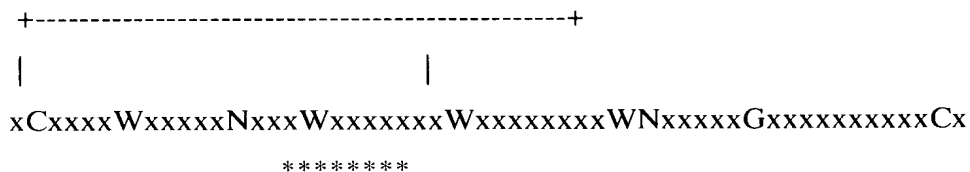
30 The microbial degradation of cellulose and xylans requires several types of enzyme such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1].

Structurally, cellulases and xylanases generally consist of a catalytic domain joined to a cellulose-binding domain (CBD) by a short linker sequence rich in proline and/or hydroxy-amino acids.

The CBD of a number of bacterial cellulases has been shown to consist of about 105 amino acid residues [2]. Enzymes known to contain such a domain are:

- Endoglucanase (gene end1) from *Butyrivibrio fibrisolvens*.
- Endoglucanases A (gene cenA) and B (cenB) from *Cellulomonas fimi*.
- Exoglucanases A (gene cbhA) and B (cbhB) from *Cellulomonas fimi*.
- Endoglucanase E-2 (gene celB) from *Thermomonospora fusca*.
- Endoglucanase A (gene celA) from *Microbispora bispora*.
- Endoglucanases A (gene celA), B (celB) and C (celC) from *Pseudomonas fluorescens*.
- Endoglucanase A (gene celA) from *Streptomyces lividans*.
- Exocellobiohydrolase (gene cex) from *Cellulomonas fimi*.
- Xylanases A (gene xynA) and B (xynB) from *Pseudomonas fluorescens*.
- Arabinofuranosidase C (EC 3.2.1.55) (xylanase C) (gene xynC) from *Pseudomonas fluorescens*.
- Chitinase 63 (EC 3.2.1.14) from *Streptomyces plicatus*.
- Chitinase C from *Streptomyces lividans*.

The CBD domain is found either at the N-terminal or at the C-terminal extremity of these enzymes. As it is shown in the following schematic representation, there are two conserved cysteines in this CBD domain - one at each extremity of the domain - which have been shown [3] to be involved in a disulfide bond. There are also four conserved tryptophan residues which could be involved in the interaction of the CBD with polysaccharides.



'C': conserved cysteine involved in a disulfide bond. '*': position of the pattern.

Consensus pattern W-N-[STAGR]-[STDN]-[LIVM]-x(2)-[GST]-x-[GST]-x(2)-[LIVMFT]-[GA]

[1] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[2] Meinke A., Gilkes N.R., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Protein Seq. Data Anal. 4:349-353(1991).

[3] Gilkes N.R., Claeysens M., Aebersold R., Henrissat B., Meinke A., Morrison H.D., Kilburn D.G., Warren R.A.J., Miller R.C. Jr. Eur. J. Biochem. 202:367-377(1991).

785. Amidases signature

It has been shown [1,2,3] that several enzymes from various prokaryotic and eukaryotic organisms which are involved in the hydrolysis of amides (amidases) are evolutionary related. These enzymes are listed below.

- Indoleacetamide hydrolase (EC 3.5.1.-), a bacterial plasmid-encoded enzyme that catalyzes the hydrolysis of indole-3-acetamide (IAM) into indole-3-acetate (IAA), the second step in the biosynthesis of auxins from tryptophan.

- Acetamidase from *Emericella nidulans* (gene *amdS*), an enzyme which allows acetamide to be used as a sole carbon or nitrogen source.

- Amidase (EC 3.5.1.4) from *Rhodococcus* sp. N-774 and *Brevibacterium* sp. R312 (gene *amdA*). This enzyme hydrolyzes propionamides efficiently, and also at a lower efficiency, acetamide, acrylamide and indoleacetamide.

- Amidase (EC 3.5.1.4) from *Pseudomonas chlororaphis*.

- 6-aminohexanoate-cyclic-dimer hydrolase (EC 3.5.2.12) (nylon oligomers degrading enzyme E1) (gene *nylA*), a bacterial plasmid encoded enzyme which catalyzes the first step in the degradation of 6-aminohexanoic acid cyclic dimer, a by-product of nylon manufacture [4].

- Glutamyl-tRNA(Gln) amidotransferase subunit A [5].

- Mammalian fatty acid amide hydrolase (gene *FAAH*) [6].

- A putative amidase from yeast (gene *AMD2*).

- *Mycobacterium tuberculosis* putative amidases *amiA2*, *amiB2*, *amiC* and *amiD*.

All these enzymes contain in their central section a highly conserved region rich in glycine, serine, and alanine residues. This region has been used as a signature pattern.

Consensus pattern: G-[GA]-S-[GS]-[GS]-G-x-[GSA]-[GSAVY]-x-[LIVM]-[GSA]-x(6)-[GSAT]-x-[GA]-x-[DE]-x-[GA]-x-S-[LIVM]-R-x-P-[GSAC]

[1] Mayaux J.-F., Cerbelaud E., Soubrier F., Faucher D., Petre D. J. Bacteriol. 172:6764-6773(1990).

[2] Hashimoto Y., Nishiyama M., Ikehata O., Horinouchi S., Beppu T. Biochim. Biophys. Acta 1088:225-233(1991).

[3] Chang T.-H., Abelson J. Nucleic Acids Res. 18:7180-7180(1990).

[4] Tsuchiya K., Fukuyama S., Kanzaki N., Kanagawa K., Negoro S., Okada H. J. Bacteriol. 171:3187-3191(1989).

[5] Curnow A.W., Hong K.W., Yuan R., Kim S.I., Martins O., Winkler W., Henkin T.M., Soll D. Proc. Natl. Acad. Sci. U.S.A. 94:11819-11826(1997).

[6] Cravatt B.F., Giang D.K., Mayfield S.P., Boger D.L., Lerner R.A., Gilula N.B. Nature 384:83-87(1996).

786. Glycosyl hydrolases family 10 active site

The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family F [3] or as the glycosyl hydrolases family 10 [4,E1]. The enzymes which are currently known to belong to this family are listed below.

- *Aspergillus awamori* xylanase A (xynA).
- *Bacillus* sp. strain 125 xylanase (xynA).
- *Bacillus stearothermophilus* xylanase.
- *Butyrivibrio fibrisolvens* xylanases A (xynA) and B (xynB).
- *Caldocellum saccharolyticum* bifunctional endoglucanase/exoglucanase (celB). This protein consists of two domains; it is the N-terminal domain, which has exoglucanase activity, which belongs to this family.
- *Caldocellum saccharolyticum* xylanase A (xynA).
- *Caldocellum saccharolyticum* ORF4. This hypothetical protein is encoded in the xynABC operon and is probably a xylanase.

651

- *Cellulomonas fimi* exoglucanase/xylanase (cex).
- *Clostridium stercoarium* thermostable celloxylanase.
- *Clostridium thermocellum* xylanases Y (xynY) and Z (xynZ).
- *Cryptococcus albidus* xylanase.

- 5 - *Penicillium chrysogenum* xylanase (gene xylP).
- *Pseudomonas fluorescens* xylanases A (xynA) and B (xynB).
- *Ruminococcus flavefaciens* bifunctional xylanase XYLA (xynA). This protein consists of three domains: a N-terminal xylanase catalytic domain that belongs to family 11 of glycosyl hydrolases; a central domain composed of short repeats of Gln, Asn and Trp, and a C-terminal
- 10 xylanase catalytic domain that belongs to family 10 of glycosyl hydrolases.
- *Streptomyces lividans* xylanase A (xlnA).
- *Thermoanaerobacter saccharolyticum* endoxylanase A (xynA).
- *Thermoascus aurantiacus* xylanase.
- Thermophilic bacterium Rt8.B4 xylanase (xynA).

One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [5], in the exoglucanase from *Cellulomonas fimi*, to be directly involved in glycosidic bond cleavage by acting as a nucleophile. This region has been used as a signature pattern.

Consensus pattern[GTA]-x(2)-[LIVN]-x-[IVMF]-[ST]-E-[LIY]-[DN]-[LIVMF] [E is the active site residue]

[1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).

25 [2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[3] Henrissat B., Claeysens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).

[4] Henrissat B. Biochem. J. 280:309-316(1991).

[5] Tull D., Withers S.G., Gilkes N.R., Kilburn D.G., Warren R.A.J., Aebersold R. J. Biol. Chem. 266:15621-15625(1991).

30

787. Fructose-bisphosphate aldolase class-II signatures

652

Fructose-bisphosphate aldolase (EC 4.1.2.13) [1,2] is a glycolytic enzyme that catalyzes the reversible aldol cleavage or condensation of fructose-1,6- bisphosphate into dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate. There are two classes of fructose-bisphosphate aldolases with different catalytic mechanisms. Class-II aldolases [2],
 5 mainly found in prokaryotes and fungi, are homodimeric enzymes which require a divalent metal ion – generally zinc - for their activity.

This family also includes the following proteins:

- Escherichia coli galactitol operon protein gatY which catalyzes the transformation of
 10 tagatose 1,6-bisphosphate into glycerone phosphate and D- glyceraldehyde 3-phosphate.
- Escherichia coli N-acetyl galactosamine operon protein agaY which catalyzes the same reaction as that of gatY.

As signature patterns for this class of enzyme, two conserved regions were selected. The first pattern is located in the first half of the sequence and contains two histidine residues that have been shown [4] to be involved in binding a zinc ion. The second is located in the C-terminal section and contains clustered acidic residues and glycines.

Consensus pattern[FYVMT]-x(1,3)-[LIVMH]-[APN]-[LIVM]-x(1,2)-[LIVM]-H-x-D-H-
 20 [GACH] [The two H's are zinc ligands]

Consensus pattern[LIVM]-E-x-E-[LIVM]-G-x(2)-[GM]-[GSTA]-x-E

[1] Perham R.N. Biochem. Soc. Trans. 18:185-187(1990).

[2] Marsh J.J., Lebherz H.G. Trends Biochem. Sci. 17:110-113(1992).

25 [3] von der Osten C.H., Barbas C.F. III, Wong C.-H., Sinskey A.J. Mol. Microbiol. 3:1625-1637(1989).

[4] Berry A., Marshall K.E. FEBS Lett. 318:11-16(1993).

788. Prolyl oligopeptidase family serine active site

30 The prolyl oligopeptidase family [1,2,3] consist of a number of evolutionary related peptidases whose catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. The known members of this family are listed below.

653

- Prolyl endopeptidase (EC 3.4.21.26) (PE) (also called post-proline cleaving enzyme). PE is an enzyme that cleaves peptide bonds on the C-terminal side of prolyl residues. The sequence of PE has been obtained from a mammalian species (pig) and from bacteria (*Flavobacterium meningosepticum* and *Aeromonas hydrophila*); there is a high degree of sequence conservation between these sequences.

- *Escherichia coli* protease II (EC 3.4.21.83) (oligopeptidase B) (gene prtB) which cleaves peptide bonds on the C-terminal side of lysyl and arginyl residues.

- Dipeptidyl peptidase IV (EC 3.4.14.5) (DPP IV). DPP IV is an enzyme that removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline.

- Yeast vacuolar dipeptidyl aminopeptidase A (DPAP A) (gene: STE13) which is responsible for the proteolytic maturation of the alpha-factor precursor.

- Yeast vacuolar dipeptidyl aminopeptidase B (DPAP B) (gene: DAP2).

- Acylamino-acid-releasing enzyme (EC 3.4.19.1) (acyl-peptide hydrolase). This enzyme catalyzes the hydrolysis of the amino-terminal peptide bond of an N-acetylated protein to generate a N-acetylated amino acid and a protein with a free amino-terminus.

A conserved serine residue has experimentally been shown (in *E. coli* protease II as well as in pig and bacterial PE) to be necessary for the catalytic mechanism. This serine, which is part of the catalytic triad (Ser, His, Asp), is generally located about 150 residues away from the C-terminal extremity of these enzymes (which are all proteins that contains about 700 to 800 amino acids).

Consensus pattern D-x(3)-A-x(3)-[LIVMFYW]-x(14)-G-x-S-x-G-G-[LIVMFYW](2) [S is the active site residue]

Note these proteins belong to families S9A/S9B/S9C in the classification of peptidases [4,E1].

[1] Rawlings N.D., Polgar L., Barrett A.J. *Biochem. J.* 279:907-911(1991).

[2] Barrett A.J., Rawlings N.D. *Biol. Chem. Hoppe-Seyler* 373:353-360(1992).

[3] Polgar L., Szabo E.

Biol. Chem. Hoppe-Seyler 373:361-366(1992).

[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

789. Formate--tetrahydrofolate ligase signatures

Formate--tetrahydrofolate ligase (EC 6.3.4.3) (formyltetrahydrofolate synthetase) (FTHFS) is one of the enzymes participating in the transfer of one-carbon units, an essential element of various biosynthetic pathways. In many of these processes the transfers of one-carbon units are mediated by the coenzyme tetrahydrofolate (THF). Various reactions generate one-carbon derivatives of THF which can be interconverted between different oxidation states by FTHFS, methylenetetrahydrofolate dehydrogenase (EC 1.5.1.5) and methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9).

In eukaryotes the FTHFS activity is expressed by a multifunctional enzyme, C-1-tetrahydrofolate synthase (C1-THF synthase), which also catalyzes the dehydrogenase and cyclohydrolase activities. Two forms of C1-THF synthases are known [1], one is located in the mitochondrial matrix, while the second one is cytoplasmic. In both forms the FTHFS domain consist of about 600 amino acid residues and is located in the C-terminal section of C1-THF synthase. In prokaryotes FTHFS activity is expressed by a monofunctional homotetrameric enzyme of about 560 amino acid residues [2].

The sequence of FTHFS is highly conserved in all forms of the enzyme. As signature patterns, two regions that are almost perfectly conserved were selected. The first one is a glycine-rich segment located in the N-terminal part of FTHFS and which could be part of an ATP-binding domain [2]. The second pattern is located in the central section of FTHFS.

Consensus pattern G-[LIVM]-K-G-G-A-A-G-G-G-Y

Consensus pattern V-A-T-[IV]-R-A-L-K-x-[HN]-G-G

[1] Shannon K.W., Rabinowitz J.C. J. Biol. Chem. 263:7717-7725(1988).

[2] Lovell C.R., Przybyla A., Ljungdahl L.G. Biochemistry 29:5687-5694(1990).

790. Transthyretin signatures

Transthyretin (prealbumin) [1] is a thyroid hormone-binding protein that seems to transport thyroxine (T4) from the bloodstream to the brain. It is a protein of about 130 amino acids that assembles as a homotetramer and forms an internal channel that binds thyroxine.

Transthyretin is mainly synthesized in the brain choroid plexus. In humans, variants of the protein are associated with distinct forms of amyloidosis.

The sequence of transthyretin is highly conserved in vertebrates. A number of uncharacterized proteins also belong to this family:

- 5 - *Escherichia coli* hypothetical protein yedX.
- *Bacillus subtilis* hypothetical protein yunM.
- *Caenorhabditis elegans* hypothetical protein R09H10.3.
- *Caenorhabditis elegans* hypothetical protein ZK697.8.

- 10 Two regions were selected as signature patterns. The first located in the N-terminal extremity starts with a lysine known to be involved in binding T4. The second pattern is located in the C-terminal extremity.

Consensus pattern[KH]-[IV]-L-[DN]-x(3)-G-x-P-A-x(2)-[IV]-x-[IV] [The K binds thyroxine]

15 Consensus patternY-[TH]-[IV]-[AP]-x(2)-L-S-[PQ]-[FYW]-[GS]-[FY]-[QS]

[1] Schreiber G., Richardson S.J. Comp. Biochem. Physiol. 116B:137-160(1997).

791. Dihydropteroate synthase signatures

20 All organisms require reduced folate cofactors for the synthesis of a variety of metabolites. Most microorganisms must synthesize folate de novo because they lack the active transport system of higher vertebrate cells which allows these organisms to use dietary folates. Enzymes that are involved in the biosynthesis of folates are therefore the target of a variety of antimicrobial agents such as trimethoprim or sulfonamides.

- 25 Dihydropteroate synthase (EC 2.5.1.15) (DHPS) catalyzes the condensation of 6-hydroxymethyl-7,8-dihydropteridine pyrophosphate to para-aminobenzoic acid to form 7,8-dihydropteroate. This is the second step in the three steps pathway leading from 6-hydroxymethyl-7,8-dihydropterin to 7,8-dihydrofolate. DHPS is the target of sulfonamides which are substrates analog that compete with para-aminobenzoic acid.

- 30 Bacterial DHPS (gene *sul* or *folP*) [1] is a protein of about 275 to 315 amino acid residues which is either chromosomally encoded or found on various antibiotic resistance plasmids. In the lower eukaryote *Pneumocystis carinii*, DHPS is the C-terminal domain of a multifunctional folate synthesis enzyme (gene *fas*) [2].

Two signature patterns for DHPS were developed, the first signature is located in the N-terminal section of these enzymes, while the second signature is located in the central section.

- 5 Consensus pattern[LIVM]-x-[AG]-[LIVMF](2)-N-x-T-x-D-S-F-x-D-x-[SG]
Consensus pattern[GE]-[SA]-x-[LIVM](2)-D-[LIVM]-G-[GP]-x(2)-[STA]-x-P

[1] Slock J., Stahly D.P., Han C.-Y., Six E.W., Crawford I.P. J. Bacteriol. 172:7211-7226(1990).

- 10 [2] Volpes F., Dyer M., Scaife J.G., Darby G., Stammers D.K., Delves C.J. Gene 112:213-218(1992).

792. Phosphatidylinositol 3- and 4-kinases signatures

Phosphatidylinositol 3-kinase (PI3-kinase) (EC 2.7.1.137) [1] is an enzyme that phosphorylates phosphoinositides on the 3-hydroxyl group of the inositol ring. The exact function of the three products of PI3-kinase - PI-3-P, PI-3,4-P(2) and PI-3,4,5-P(3) - is not yet known, although it is proposed that they function as second messengers in cell signalling. Currently, three forms of PI3-kinase are known:

- The mammalian enzyme which is a heterodimer of a 110 Kd catalytic chain (p110) and an 85 Kd subunit (p85) which allows it to bind to activated tyrosine protein kinases. There are at least two different types of p100 subunits (alpha and beta).
- Yeast TOR1/DRR1 and TOR2/DRR2 [2], PI3-kinases required for cell cycle activation. Both are proteins of about 280 Kd.
- Yeast VPS34 [3], a PI3-kinase involved in vacuolar sorting and segregation. VPS34 is a protein of about 100 Kd.
- Arabidopsis thaliana and soybean VPS34 homologs.

Phosphatidylinositol 4-kinase (PI4-kinase) (EC 2.7.1.67) [4] is an enzyme that acts on phosphatidylinositol (PI) in the first committed step in the production of the second messenger inositol-1,4,5,-trisphosphate. Currently the following forms of PI4-kinases are known:

- Human PI4-kinase alpha.
- Yeast PIK1, a nuclear protein of 120 Kd.

- Yeast STT4, a protein of 214 Kd.

The PI3- and PI4-kinases share a well conserved domain at their C-terminal section; this domain seems to be distantly related to the catalytic domain of protein kinases [2]. Two signature patterns were developed from the best conserved parts of this domain.

Four additional proteins belong to this family:

- Mammalian FKBP-rapamycin associated protein (FRAP) [5], which acts as the target for the cell-cycle arrest and immunosuppressive effects of the FKBP12-rapamycin complex.

- Yeast protein ESR1 [6] which is required for cell growth, DNA repair and meiotic recombination.

- Yeast protein TEL1 which is involved in controlling telomere length.

- Yeast hypothetical protein YHR099w, a distantly related member of this family.

- Fission yeast hypothetical protein SpAC22E12.16C.

Consensus pattern[LIVMFAC]-K-x(1,3)-[DEA]-[DE]-[LIVMC]-R-Q-[DE]-x(4)-Q

Consensus pattern[GS]-x-[AV]-x(3)-[LIVM]-x(2)-[FYH]-[LIVM](2)-x-[LIVMF]-x-D-R-H-x(2)-N

[1] Hiles I.D., Otsu M., Volinia S., Fry M.J., Gout I., Dhand R., Panayotou G., Ruiz-Larrea F., Thompson A., Totty N.F., Hsuan J.J., Courtneidge S.A., Parker P.J., Waterfield M.D. Cell 70:419-429(1992).

[2] Kunz J., Henriquez R., Schneider U., Deuter-Reinhard M., Movva N., Hall M.N. Cell 73:585-596(1993).

[3] Schu P.V., Takegawa K., Fry M.J., Stack J.H., Waterfield M.D., Emr S.D. Science 260:88-91(1993).

[4] Garcia-Bustos J.F., Marini F., Stevenson I., Frei C., Hall M.N. EMBO J. 13:2352-2361(1994).

[5] Brown E.J., Albers M.W., Shin T.B., Ichikawa K., Keith C.T., Lane W.S., Schreiber S.L. Nature 369:756-758(1994).

[6] Kato R., Ogawa H. Nucleic Acids Res. 22:3104-3112(1994).

793. FAD-dependent glycerol-3-phosphate dehydrogenase signatures

FAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) (GPD) catalyzes the conversion of glycerol-3-phosphate into dihydroxyacetone phosphate. In bacteria [1] it is associated with the utilization of glycerol coupled to respiration. In *Escherichia coli*, two isozymes are known: one expressed under anaerobic conditions (gene *glpA*) and one in aerobic conditions (gene *glpD*). In eukaryotes, a mitochondrial form of GPD participates in the glycerol phosphate shuttle in conjunction with an NAD-dependent cytoplasmic GPD (EC 1.1.1.8) [2,3].

These enzymes are proteins of about 60 to 70 Kd which contain a probable FAD-binding domain in their N-terminal extremity. The mammalian enzyme differs from the bacterial or yeast proteins by having an EF-hand calcium-binding region (See <PDOC00018>) in its C-terminal extremity.

Two signature patterns were developed. One based on the first half of the FAD-binding domain and one which corresponds to a conserved region in the central part of these enzymes.

Consensus pattern[IV]-G-G-G-x(2)-G-[STACV]-G-x-A-x-D-x(3)-R-G

Consensus patternG-G-K-x(2)-[GSTE]-Y-R-x(2)-A

[1] Austin D., Larson T.J. J. Bacteriol. 173:101-107(1991).

[2] Roennow B., Kielland-Brandt M.C. Yeast 9:1121-1130(1993).

[3] Brown L.J., McDonald M.J., Lehn D.A., Moran S.M. J. Biol. Chem. 269:14363-14366(1994).

794. NOL1/NOP2/sun family signature

The following proteins seems to be evolutionary related:

- Mammalian proliferating-cell nucleolar antigen p120 (gene *NOL1*) which may play a role in the regulation of the cell cycle and the increased nucleolar activity that is associated with the cell proliferation.

- Yeast nucleolar protein NOP2 (or *YNA1*) which could be involved in nucleolar function during the onset of growth, and in the maintenance of nucleolar structure.

- Yeast hypothetical protein YBL024w.

- Bacterial protein sun (also known as *fmU*).

- *Escherichia coli* hypothetical protein yebU.

659

- Mycobacterium tuberculosis hypothetical protein MtCY21B4.24.
- Methanococcus jannaschii hypothetical protein MJ0026.

NOL1 is a protein of 855 residues, NOP2 consists of 618 residues, YBL024w of 684, sun is a protein of about 430 to 450 residues and MJ026 has 274 residues. They share a conserved central domain which contains some highly conserved regions. One of these regions was selected as a signature pattern.

Consensus pattern[FV]-D-[KRA]-[LIVMA]-L-x-D-[AV]-P-C-[ST]-[GA]

795. moaA / nifB / pqqE family signature

A number of proteins involved in the biosynthesis of metallo cofactors have been shown [1,2] to be evolutionary related. These proteins are:

- Bacterial and archebacterial protein moaA, which is involved in the biosynthesis of the molybdenum cofactor (molybdopterin; MPT).
- Arabidopsis thaliana cnx2, a protein involved in molybdopterin biosynthesis and which is highly similar to moaA.
- Bacillus subtilis narA, which seems to be the moaA ortholog in that bacteria.
- Bacterial protein nifB (or fixZ) which is involved in the biosynthesis of the nitrogenase iron-molybdenum cofactor.
- Bacterial protein pqqE which is involved in the biosynthesis of the cofactor pyrrolo-quinoline-quinone (PQQ).
- Pyrococcus furiosus cmo, a protein involved in the synthesis of a molybdopterin-based tungsten cofactor.
- Caenorhabditis elegans hypothetical protein F49E2.1.

All these proteins share, in their N-terminal region, a conserved domain that contains three cysteines. In moaA, these cysteines have been shown [1] to be important for the biological activity. They could be involved in the binding of an iron-sulfur cluster.

Consensus pattern[LIV]-x(3)-C-[NP]-[LIVMF]-[QRS]-C-x-[FYM]-C [The three C's are putative Fe-S ligands]

- [1] Menendez C., Igloi G., Henninger H., Brandsch R. Arch. Microbiol. 164:142-151(1995).
[2] Hoff T., Schnorr K.M., Meyer C., Caboche M. J. Biol. Chem. 270:6100-6107(1995).

796. Forkhead-associated (FHA) domain profile

5 The forkhead-associated (FHA) domain [1,E1] is a putative nuclear signalling domain found in a variety of otherwise unrelated proteins. The FHA domain comprise approximately 55 to 75 amino acids and contains three highly conserved blocks separated by divergent spacer regions. Currently it has been found in the following proteins:

- Four transcription factors that also contain a forkhead (FH) domain: mouse myocyte

10 nuclear factor 1 (MNF1), yeast transcription factor FHL1, which probably controls pre-mRNA processing, and yeast FKH1 and FKH2. In those protein the FHA domain is located N-terminal of the DNA-binding FH domain.

- Kinase-associated protein phosphatase (KAPP) from Arabidopsis thaliana, a protein which specifically interacts with the receptor-type Ser/Thr-kinase RLK5. In KAPP, the FHA
5 domain maps to a region that interacts with the receptor-type protein kinase RLK5 only if the kinase is phosphorylated on serine residues [2].

- Two protein kinases from yeast that are involved in mediating the nuclear response to DNA damage: DUN1 and SPK1/SAD1 [3]. The latter is the only known protein containing two
10 copies of the FHA domain.

20 - Protein kinase cds1 from fission yeast contains a FHA domain and might be the ortholog of SPK1.

- Protein kinase MEK1 from yeast, which is involved in meiotic recombination.

- Human nuclear antigen Ki67 which is expressed only in proliferating cells.

25 - Yeast hypothetical protein YHR115c, which contains a RING-finger C-terminal of the FHA domain.

- Yeast hypothetical proteins L8083.1 and 9346.10, which contain an extensive coiled-coil region C-terminal of the FHA domain.

- Caenorhabditis elegans hypothetical protein ZK632.2.

- Caenorhabditis elegans hypothetical protein C01G6.5.

30 - FraH from the prokaryote Anabaena, which contains a zinc-finger motif N-terminal of the FHA domain.

- An ORF from the bacterium Streptomyces, which is on the opposite strand of the protein kinase pks1, overlapping the ORF of the kinase.

[1] Hofmann K.O., Bucher P. Trends Biochem. Sci. 20:347-349(1995).

[2] Stone J.M., Collinge M.A., Smith R.D., Horn M.A., Walker J.C. Science 266:793-795(1994).

5 [3] Navas T.A., Zhou Z., Elledge S.J. Cell 80:29-39(1995).

797. Ald_Xan_dh_C

Aldehyde oxidase and xanthine dehydrogenase, C terminus

10 [1] Romao MJ, Archer M, Moura I, Moura JJ, LeGall J, Engh R, Schneider M, Hof P, Huber R; Medline: 96072968 "Crystal structure of the xanthine oxidase-related aldehyde oxidoreductase from D. gigas." Science 1995;270:1170-1176.

Number of members: 54

5 798. Glyco_hydro_38

Glycosyl hydrolases family 38

Glycosyl hydrolases are key enzymes of carbohydrate metabolism.

20 Number of members: 20

[1] Henrissat B; Medline: 98313424; "Glycosidase families" Biochem Soc Trans 1998;26:153-156.

25 799. HECT

HECT-domain (ubiquitin-transferase).

The name HECT comes from Homologous to the E6-AP Carboxyl Terminus.

30 Number of members: 43

[1] Huibregtse JM, Scheffner M, Beaudenon S, Howley PM; Medline: 95223981; "A family of proteins structurally and functionally related to the E6-AP ubiquitin-protein ligase." Proc Natl Acad Sci U S A 1995;92:2563-2567.

5 800. HRDC

HRDC domain

The HRDC (Helicase and RNase D C-terminal) domain has a putative role in nucleic acid binding. Mutations in the HRDC domain cause human disease.

10 Number of members: 19

[1] Morozov V, Mushegian AR, Koonin EV, Bork P; Medline: 98060076; "A putative nucleic acid-binding domain in Bloom's and Werner's syndrome helicases" Trends Biochem Sci 1997;22:417-418.

15 801. Integrase

Integrase mediates integration of a DNA copy of the viral genome into the host chromosome. Integrase is composed of three domains. The amino-terminal domain is a zinc binding domain. The central domain is the catalytic domain [1]. The carboxyl terminal domain is a DNA binding domain [2].

20 Number of members: 581

[1] Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR; Medline: 95099322. "Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases." Science 1994;266:1981-1986.

[2] Lodi PJ, Ernst JA, Kuszewski J, Hickman AB, Engelman A, Craigie R, Clore GM, Gronenborn AM; Medline: 95359147; "Solution structure of the DNA binding domain of HIV-1 integrase." Biochemistry 1995;34:9826-9833

30

802. lig_chan

Ligand-gated ion channel

This family includes the four transmembrane regions of the ionotropic glutamate receptors and NMDA receptors.

Number of members: 128

5

[1] Tong G, Shepherd D, Jahr CE; Medline: 95184014; "Synaptic desensitization of NMDA receptors by calcineurin." Science 1995;267:1510-1512.

803. RhoGAP

10 RhoGAP domain

GTPase activator proteins towards Rho/Rac/Cdc42-like small GTPases.

Number of members: 97

15

[1] Musacchio A, Cantley LC, Harrison SC; Medline: 97121392; "Crystal structure of the breakpoint cluster region-homology domain from phosphoinositide 3-kinase p85 alpha subunit." Proc Natl Acad Sci U S A 1996;93:14373-14378.

20

[2] Barrett T, Xiao B, Dodson EJ, Dodson G, Ludbrook SB, Nurmahomed K, Gamblin SJ, Musacchio A, Smerdon SJ, Eccleston JF; Medline: 97162209; "The structure of the GTPase-activating domain from p50rhoGAP." Nature 1997;385:458-461.

[3] Rittinger K, Walker PA, Eccleston JF, Nurmahomed K, Owen D, Laue E, Gamblin SJ, Smerdon SJ; Medline: 97404320; "Crystal structure of a small G protein in complex with the GTPase-activating protein rhoGAP." Nature 1997;388:693-697.

25

[4] Boguski MS, McCormick F; Medline: 94081948; "Proteins regulating Ras and its relatives." Nature 1993;366:643-654.

804. vwd

von Willebrand factor type D domain

30

[1] Bork P; Medline: 93327926; "The modular architecture of a new family of growth regulators related to connective tissue growth factor." FEBS lett 1993;327:125-130.

Number of members: 92

805. zf-C4_Topoisom

Topoisomerase DNA binding C4 zinc finger

- 5 [1] Tse-Dinh YC, Beran-Steed RK; Medline: 89034032; "Escherichia coli DNA topoisomerase I is a zinc metalloprotein with three repetitive zinc-binding domains." J Biol Chem 1988;263:15857-15859.
- [2] Ahumada A, Tse-Dinh YC; Medline: 99011409; "The Zn(II) binding motifs of E. coli DNA topoisomerase I is part of a high-affinity DNA binding domain." Biochem Biophys Res Commun 1998;251:509-514.
- 10

Number of members: 51

15 806. AIRC

AIR carboxylase

Members of this family catalyse the decarboxylation of 1-(5-phosphoribosyl)-5-amino-4-imidazole-carboxylate (AIR). This family catalyse the sixth step of de novo purine biosynthesis. Some members of this family contain two copies of this domain. Number of members: 35

20

807. Bromodomain signature and profile

PROSITE cross-reference(s): PS00633; BROMODOMAIN_1, PS50014; BROMODOMAIN_2

- 25 The bromodomain [1,2,3] is a conserved region of about 70 amino acids found in the following proteins:

- Higher eukaryotes transcription initiation factor TFIID 250 Kd subunit (TBP-associated factor p250) (gene CCG1). P250 associated with the TFIID TATA-box binding protein and seems essential for progression of the G1 phase of the cell cycle.
 - Human RING3, a protein of unknown function encoded in the MHC class II locus.
 - Mammalian CREB-binding protein (CBP), which mediates cAMP-gene regulation by binding specifically to phosphorylated CREB protein.
- 30

- *Drosophila* female sterile homeotic protein (gene *fsh*), required maternally for proper expression of other homeotic genes involved in pattern formation, such as *Ubx*.

- *Drosophila* *brahma* protein (gene *brm*), a protein required for the activation of multiple homeotic genes.

5 - Mammalian homologs of *brahma*. In human, three *brahma*-like proteins are known: SNF2a(hBRM), SNF2b, and BRG1.

- Human BS69, a protein that binds to adenovirus E1A and inhibits E1A transactivation

- Human peregrin (or Br140).

10 - Yeast BDF1 [3], a transcription factor involved in the expression of a broad class of genes including snRNAs.

- Yeast GCN5, a general transcriptional activator operating in concert with certain other DNA-binding transcriptional activators, such as GCN4, HAP2/3/4 or ADA2.

- Yeast NPS1/STH1, involved in G(2) phase control in mitosis.

15 - Yeast SNF2/SWI2, which is part of a complex with the SNF5, SNF6, SWI3 and ADR6/SWI1 proteins. This SWI-complex is involved in transcriptional activation.

- Yeast SPT7, a transcriptional activator of Ty elements and possibly other genes.

- *Caenorhabditis elegans* protein *cbp-1*.

- Yeast hypothetical protein YGR056w.

- Yeast hypothetical protein YKR008w.

20 - Yeast hypothetical protein L9638.1.

Some proteins contain a region which, while similar to some extent to a classical bromodomain, diverges from it by either lacking part of the domain or because of an insertion. These proteins are:

25

- Mammalian protein HRX (also known as All-1 or MLL), a protein involved in translocations leading to acute leukemias and which possibly acts as a transcriptional regulatory factor. HRX contains a region similar to the C- terminal half of the bromodomain.

30 - *Caenorhabditis elegans* hypothetical protein ZK783.4. The bromodomain of this protein has a 23 amino-acid insertion.

- Yeast protein YTA7. This protein contains a region with significant similarity to the C-terminal half of the bromodomain. As it is a member of the AAA family (see <PDOC00572>) it is also in a functionally different context.

The above proteins generally contain a single bromodomain, but some of them contain two copies, this is the case of BDF1, CCG1, fsh, RING3, YKR008w and L9638.1.

- 5 The exact function of this domain is not yet known but it is thought to be involved in protein-protein interactions and it may be important for the assembly or activity of multicomponent complexes involved in transcriptional activation.

10 The consensus pattern that has been developed spans a major part of the bromodomain; a more sensitive detection is available through the use of a profile which spans the whole domain.

Consensus pattern[STANVF]-x(2)-F-x(4)-[DNS]-x(5,7)-[DENQTF]-Y-[HFY]-x(2)-
[LIVMFY]-x(3)-[LIVM]-x(4)-[LIVM]-x(6,8)-Y-x(12,13)-[LIVM]-
15 x(2)-N-[SACF]-x(2)-[FY]

References

- 20 [1] Haynes S.R., Doolard C., Winston F., Beck S., Trowsdale J., Dawid I.B. Nucleic Acids Res. 20:2693-2603(1992).
[2] Tamkun J.W., Deuring R., Scott M.P., Kissinger M., Pattatucci A.M., Kaufman T.C., Kennison J.A. Cell 68:561-572(1992).
[3] Tamkun J.W. Curr. Opin. Genet. Dev. 5:473-477(1995).

808. (CH) Actinin-type actin-binding domain signatures

25 PROSITE cross-reference(s): PS00019; ACTININ_1, PS00020; ACTININ_2

Alpha-actinin is a F-actin cross-linking protein which is thought to anchor actin to a variety of intracellular structures [1]. The actin-binding domain of alpha-actinin seems to reside in the first 250 residues of the protein. A similar actin-binding domain has been found in the N-
30 terminal region of many different actin-binding proteins [2,3]:

- In the beta chain of spectrin (or fodrin).

- In dystrophin, the protein defective in Duchenne muscular dystrophy (DMD) and which may play a role in anchoring the cytoskeleton to the plasma membrane.

- In the slime mold gelation factor (or ABP-120).

- In actin-binding protein ABP-280 (or filamin), a protein that link actin filaments to membrane glycoproteins.

- In fimbrin (or plastin), an actin-bundling protein. Fimbrin differs from the above proteins in that it contains two tandem copies of the actin-binding domain and that these copies are located in the C-terminal part of the protein.

Two conserved regions were selected as signature patterns for this type of main. The first of this region is located at the beginning of the domain, hile the second one is located in the central section and has been shown to be essential for the binding of actin.

Consensus pattern[EQ]-x(2)-[ATV]-[FY]-x(2)-W-x-N

Consensus pattern[LIVM]-x-[SGN]-[LIVM]-[DAGHE]-[SAG]-x-[DNEAG]-[LIVM]-x-[DEAG]-x(4)-[LIVM]-x-[LM]-[SAG]-[LIVM]-[LIVMT]-W-x- [LIVM](2)

[1] Schleicher M., Andre E., Harmann A., Noegel A.A. Dev. Genet. 9:521-530(1988).

[2] Matsudaira P. Trends Biochem. Sci. 16:87-92(1991).

[3] Dubreuil R.R. BioEssays 13:219-226(1991).

809. (COX1) Heme-copper oxidase subunit I, copper B binding region signature
PROSITE cross-reference(s): PS00077; COX1

Heme-copper respiratory oxidases [1] are oligomeric integral membrane protein complexes that catalyze the terminal step in the respiratory chain: they transfer electrons from cytochrome c or a quinol to oxygen. Some terminal oxidases generate a transmembrane proton gradient across the plasma membrane (prokaryotes) or the mitochondrial inner membrane (eukaryotes). The enzyme complex consists of 3-4 subunits (prokaryotes) up to 13 polypeptides (mammals) of which only the catalytic subunit (equivalent to mammalian subunit 1 (CO I)) is found in all heme-copper respiratory oxidases. The presence of a bimetallic center (formed by a high-spin heme and copper B) as well as a low-spin heme, both ligated to six conserved histidine residues near the outer side of four

transmembrane spans within CO I is common to all family members [2-4].

In contrary to eukaryotes the respiratory chain of prokaryotes is branched to multiple terminal oxidases. The enzyme complexes vary in heme and copper composition, substrate type and substrate affinity. The different respiratory oxidases allow the cells to customize their respiratory systems according a variety of environmental growth conditions [1].

Recently also a component of an anaerobic respiratory chain has been found to contain the copper B binding signature of this family: nitric oxide reductase (NOR) exists in denitrifying species of Archae and Eubacteria.

Enzymes that belong to this family are:

- Mitochondrial-type cytochrome c oxidase (EC 1.9.3.1) which uses cytochrome c as electron donor. The electrons are transferred via copper A (Cu(A)) and heme a to the bimetallic center of CO I that is formed by a penta-coordinated heme a and copper B (Cu(B)). Subunit 1 contains 12 transmembrane regions. Cu(B) is said to be ligated to three of the conserved histidine residues within the transmembrane segments 6 and 7.
- Quinol oxidase from prokaryotes that transfers electrons from a quinol to the binuclear center of polypeptide I. This category of enzymes includes Escherichia coli cytochrome O terminal oxidase complex which is a component of the aerobic respiratory chain that predominates when cells are grown at high aeration.
- FixN, the catalytic subunit of a cytochrome c oxidase expressed in nitrogen-fixing bacteroids living in root nodules. The high affinity for oxygen allows oxidative phosphorylation under low oxygen concentrations. A similar enzyme has been found in other purple bacteria.
- Nitric oxide reductase (EC 1.7.99.7) from Pseudomonas stutzeri. NOR reduces nitrate to dinitrogen. It is a heterodimer of norC and the catalytic subunit norB. The latter contains the 6 invariant histidine residues and 12 transmembrane segments [5].

As a signature pattern the copper-binding region was used.

Consensus pattern[YWG]-[LIVFYWTA](2)-[VGS]-H-[LNP]-x-V-x(44,47)-H-H [The
5 three H's are copper B ligands]

Notocytochrome bd complexes do not belong to this family.

[1]

10 Garcia-Horsman J.A., Barquera B., Rumbley J., Ma J., Gennis R.B.
J. Bacteriol. 176:5587-5600(1994).

[2]

Castresana J., Luebben M., Saraste M., Higgins D.G.
EMBO J. 13:2516-2525(1994).

[3]

Capaldi R.A., Malatesta F., Darley-Usmar V.M.
Biochim. Biophys. Acta 726:135-148(1983).

[4]

Holm L., Saraste M., Wikstrom M.
EMBO J. 6:2819-2823(1987).

[5]

Saraste M., Castresana J.
FEBS Lett. 341:1-4(1994).

25 810. (dehydrog_molyb) Eukaryotic molybdopterin oxidoreductases signature
PROSITE cross-reference(s): PS00559; MOLYBDOPTERIN_EUK

A number of different eukaryotic oxidoreductases that require and bind a
molybdopterin cofactor have been shown [1] to share a few regions of sequence
30 similarity. These enzymes are:

- Xanthine dehydrogenase (EC 1.1.1.204), which catalyzes the oxidation of
xanthine to uric acid with the concomitant reduction of NAD. Structurally,

this enzyme of about 1300 amino acids consists of at least three distinct domains: an N-terminal 2Fe-2S ferredoxin-like iron-sulfur binding domain (see <PDOC00175>), a central FAD/NAD-binding domain and a C-terminal Mo-pterin domain.

5 - Aldehyde oxidase (EC 1.2.3.1), which catalyzes the oxidation aldehydes into acids. Aldehyde oxidase is highly similar to xanthine dehydrogenase in its sequence and domain structure.

- Nitrate reductase (EC 1.6.6.1), which catalyzes the reduction of nitrate to nitrite. Structurally, this enzyme of about 900 amino acids consists of
10 an N-terminal Mo-pterin domain, a central cytochrome b5-type heme-binding domain (see <PDOC00170>) and a C-terminal FAD/NAD-binding cytochrome reductase domain.

- Sulfite oxidase (EC 1.8.3.1), which catalyzes the oxidation of sulfite to sulfate. Structurally, this enzyme of about 460 amino acids consists of an
15 N-terminal cytochrome b5-binding domain followed by a Mo-pterin domain.

There are a few conserved regions in the sequence of the molybdopterin-binding domain of these enzymes. The pattern uses to detect these proteins is based
20 on one of them. It contains a cysteine residue which could be involved in binding the molybdopterin cofactor.

Consensus pattern[GA]-x(3)-[KRNQHT]-x(11,14)-[LIVMFYWS]-x(8)-[LIVMF]-x-C-x(2)-[DEN]-R-x(2)-[DE]

25 [1]

Wootton J.C., Nicolson R.E., Cock J.M., Walters D.E., Burke J.F., Doyle W.A., Bray R.C.

Biochim. Biophys. Acta 1057:157-185(1991).

30 811. (DNA_ligase) ATP-dependent DNA ligase signatures

, PROSITE cross-reference(s): PS00697; DNA_LIGASE_A1, PS00333; DNA_LIGASE_A2

DNA ligase (polydeoxyribonucleotide synthase) is the enzyme that joins two DNA

671

fragments by catalyzing the formation of an internucleotide ester bond between phosphate and deoxyribose. It is active during DNA replication, DNA repair and DNA recombination. There are two forms of DNA ligase: one requires ATP (EC 6.5.1.1), the other NAD (EC 6.5.1.2).

5

Eukaryotic, archaebacterial, virus and phage DNA ligases are ATP-dependent. During the first step of the joining reaction, the ligase interacts with ATP to form a covalent enzyme-adenylate intermediate. A conserved lysine residue is the site of adenylation [1,2].

10

Apart from the active site region, the only conserved region common to all ATP-dependent DNA ligases is found [3] in the C-terminal section and contains a conserved glutamate as well as four positions with conserved basic residues.

15

Signature patterns were developed for both conserved regions.

Consensus pattern[EDQH]-x-K-x-[DN]-G-x-R-[GACIVM] [K is the active site residue]

20

Consensus patternE-G-[LIVMA]-[LIVM](2)-[KR]-x(5,8)-[YW]-[QNEK]-x(2,6)-[KRH]-x(3,5)-K-[LIVMFY]-K

Sequences known to belong to this class detected by the patternALL, except for archebacterial DNA ligases.

25

[1]

Tomkinson A.E., Totty N.F., Ginsburg M., Lindahl T.
Proc. Natl. Acad. Sci. U.S.A. 88:400-404(1991).

[2]

Lindahl T., Barnes D.E.

30

Annu. Rev. Biochem. 61:251-281(1992).

[3]

Kletzin A.

Nucleic Acids Res. 20:5389-5396(1992).

812. (FAD_Gly3P_dh) FAD-dependent glycerol-3-phosphate dehydrogenase signatures
PROSITE cross-reference(s): PS00977; FAD_G3PDH_1, PS00978; FAD_G3PDH_2

5 FAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) (GPD) catalyzes
the conversion of glycerol-3-phosphate into dihydroxyacetone phosphate. In
bacteria [1] it is associated with the utilization of glycerol coupled to
respiration. In *Escherichia coli*, two isozymes are known: one expressed under
anaerobic conditions (gene *glpA*) and one in aerobic conditions (gene *glpD*). In
10 eukaryotes, a mitochondrial form of GPD participates in the glycerol phosphate
shuttle in conjunction with an NAD-dependent cytoplasmic GPD (EC 1.1.1.8) [2,
3].

These enzymes are proteins of about 60 to 70 Kd which contain a probable
15 FAD-binding domain in their N-terminal extremity. The mammalian enzyme differs
from the bacterial or yeast proteins by having an EF-hand calcium-binding
region (See <PDOC00018>) in its C-terminal extremity.

Two signature patterns were developed. One based on the first half of the FAD-
20 binding domain and one which corresponds to a conserved region in the central
part of these enzymes.

Consensus pattern[IV]-G-G-G-x(2)-G-[STACV]-G-x-A-x-D-x(3)-R-G

25 Consensus patternG-G-K-x(2)-[GSTE]-Y-R-x(2)-A
[1]

Austin D., Larson T.J.

J. Bacteriol. 173:101-107(1991).

[2]

30 Roennow B., Kielland-Brandt M.C.
Yeast 9:1121-1130(1993).

[3]

Brown L.J., McDonald M.J., Lehn D.A., Moran S.M.

J. Biol. Chem. 269:14363-14366(1994).

813. (Fapy_DNA_glyco) Formamidopyrimidine-DNA glycosylase signature
PROSITE cross-reference(s): PS01242; FPG

5

Formamidopyrimidine-DNA glycosylase (EC 3.2.2.23) [1] (Fapy-DNA glycosylase) (gene fpg) is a bacterial enzyme involved in DNA repair and which excise oxidized purine bases to release 2,6-diamino-4-hydroxy-5N-methylformamido-pyrimidine (Fapy) and 7,8-dihydro-8-oxoguanine (8-OxoG) residues. In addition to its glycosylase activity, FPG can also nick DNA at apurinic/apyrimidinic sites (AP sites). FPG is a monomeric protein of about 32 Kd which binds and require zinc for its activity.

10

The binding site for zinc seems to be located in the C-terminal part of the enzyme where four conserved and essential [2] cysteines are located. A signature pattern was developed based on this region.

Consensus pattern C-x(2,4)-C-x-[GTAQ]-x-[IV]-x(7)-R-[GSTAN]-[STA]-x-[FYI]-C-x(2)-C-Q

[The four C's are putative zinc ligands]

[1]

Duwat P., de Oliveira R., Ehrlich S.D., Boiteux S.
Microbiology 141:411-417(1995).

[2]

O'Connor T.E., Graves R.J., Demurcia G., Castaing B., Laval J.
J. Biol. Chem. 268:9063-9070(1993).

814. (G_glu_transpept) Gamma-glutamyltranspeptidase signature

PROSITE cross-reference(s): PS00462; G_GLU_TRANSPEPTIDASE

Gamma-glutamyltranspeptidase (EC 2.3.2.2) (GGT) [1] catalyzes the transfer of the gamma-glutamyl moiety of glutathione to an acceptor that may be an amino

674

acid, a peptide or water (forming glutamate). GGT plays a key role in the gamma-glutamyl cycle, a pathway for the synthesis and degradation of glutathione. In prokaryotes and eukaryotes, it is an enzyme that consists of two polypeptide chains, a heavy and a light subunit, processed from a single chain precursor. The active site of GGT is known to be located in the light subunit.

The sequences of mammalian and bacterial GGT show a number of regions of high similarity [2]. Pseudomonas cephalosporin acylases (EC 3.5.1.-) that convert 7-beta-(4-carboxybutanamido)-cephalosporanic acid (GL-7ACA) into 7-aminocephalosporanic acid (7ACA) and glutaric acid are evolutionary related to GGT and also show some GGT activity [3]. Like GGT, these GL-7ACA acylases, are also composed of two subunits.

One of the conserved regions correspond to the N-terminal extremity of the mature light chains of these enzymes. This region was used as a signature pattern.

Consensus pattern T-[STA]-H-x-[ST]-[LIVMA]-x(4)-G-[SN]-x-V-[STA]-x-T-x-T-[LIVM]-[NE]-x(1,2)-[FY]-G

[1]

Tate S.S., Meister A.

Meth. Enzymol. 113:400-419(1985).

[2]

Suzuki H., Kumagai H., Echigo T., Tochikura T.

J. Bacteriol. 171:5169-5172(1989).

[3]

Ishiye M., Niwa M.

Biochim. Biophys. Acta 1132:233-239(1992).

815. G-protein gamma subunit profile

PROSITE cross-reference(s): PS50058; G_PROTEIN_GAMMA

Guanine nucleotide-binding proteins (G proteins) [1] act as intermediaries in the transduction of signals generated by transmembrane receptors. G proteins consist of three subunits (alpha, beta, and gamma). The alpha subunit binds to and hydrolyzes GTP; the functions of the beta and gamma subunits are less clear but they seem to be required for the replacement of GDP by GTP as well as for membrane anchoring and receptor recognition.

The gamma subunits are small proteins (from 70 to 110 residues) that are bound to the membrane via a isoprenyl group (either a farnesyl or a geranyl-geranyl) covalently linked to their C-terminus. In mammals there are at least 12 different isoforms of gamma subunits.

The *Caenorhabditis elegans* protein egl-10, which is a regulator of G-protein signalling, contains a G-protein gamma-like domain.

A profile was developed that spans the complete length of the gamma subunit.

[1]

Pennington S.R.

Protein Prof. 2:16-315(1995).

816. GNS1/SUR4 family signature

PROSITE cross-reference(s): PS01188; GNS1_SUR4

The following group of eukaryotic integral membrane proteins, whose exact function has not yet clearly been established, are evolutionary related [1]:

- Yeast GNS1 [2], a protein involved in synthesis of 1,3-beta-glucan.
- Yeast SUR4 (or APA1, SRE1) [3], a protein that could act in a glucose-signaling pathway that controls the expression of several genes that are transcriptionally regulated by glucose.

- Yeast hypothetical protein YJL196c.
- Caenorhabditis elegans hypothetical protein C40H1.4.
- Caenorhabditis elegans hypothetical protein D2024.3.

5 The proteins have from 290 to 435 amino acid residues. Structurally, they seem to be formed of three sections: a N-terminal region with two transmembrane domains, a central hydrophilic loop and a C-terminal region that contains from one to three transmembrane domains. A conserved region that contains three histidines was selected as a signature pattern. This region is located in the
10 hydrophilic loop.

Consensus pattern L-x-F-L-H-x-Y-H-H

[1]

5 Bairoch A.

Unpublished observations (1996).

[2]

El-Sherbeini M., Clemas J.A.

J. Bacteriol. 177:3227-3234(1995).

20 [3]

Garcia-Arranz M., Maldonado A.M., Mazon M.J., Portillo F.

J. Biol. Chem. 269:18076-18082(1994).

817. Immunoglobulins and major histocompatibility complex proteins signature

25 PROSITE cross-reference(s): PS00290; IG_MHC

The basic structure of immunoglobulin (Ig) [1] molecules is a tetramer of two light chains and two heavy chains linked by disulfide bonds. There are two types of light chains: kappa and lambda, each composed of a constant domain
30 (CL) and a variable domain (VL). There are five types of heavy chains: alpha, delta, epsilon, gamma and mu, all consisting of a variable domain (VH) and three (in alpha, delta and gamma) or four (in epsilon and mu) constant domains (CH1 to CH4).

The major histocompatibility complex (MHC) molecules are made of two chains. In class I [2] the alpha chain is composed of three extracellular domains, a transmembrane region and a cytoplasmic tail. The beta chain (beta-2-microglobulin) is composed of a single extracellular domain. In class II [3], both the alpha and the beta chains are composed of two extracellular domains, a transmembrane region and a cytoplasmic tail.

It is known [4,5] that the Ig constant chain domains and a single extracellular domain in each type of MHC chains are related. These homologous domains are approximately one hundred amino acids long and include a conserved intradomain disulfide bond. A small pattern around the C-terminal cysteine is involved in this disulfide bond which can be used to detect these category of Ig related proteins.

Consensus pattern[FY]-x-C-x-[VA]-x-H-Sequences known to belong to this class detected by the pattern: Ig heavy chains type Alpha C region : All, in CH2 and CH3. Ig heavy chains type Delta C region : All, in CH3. Ig heavy chains type Epsilon C region: All, in CH1, CH3 and CH4. Ig heavy chains type Gamma C region : All, in CH3 and also CH1 in some cases Ig heavy chains type Mu C region : All, in CH2, CH3 and CH4. Ig light chains type Kappa C region : In all CL except rabbit and Xenopus. Ig light chains type Lambda C region : In all CL except rabbit. MHC class I alpha chains : All, in alpha-3 domains, including in the cytomegalovirus MHC-1 homologous protein [6]. Beta-2-microglobulin : All. MHC class II alpha chains: All, in alpha-2 domains. MHC class II beta chains: All, in beta-2 domains.

[1]

Gough N.

Trends Biochem. Sci. 6:203-205(1981).

[2]

Klein J., Figueroa F.

Immunol. Today 7:41-44(1986).

[3]

Figueroa F., Klein J.

Immunol. Today 7:78-81(1986).

[4]

5 Orr H.T., Lancet D., Robb R.J., Lopez de Castro J.A., Strominger J.L.

Nature 282:266-270(1979).

[5]

Cushley W., Owen M.J.

Immunol. Today 4:88-92(1983).

10 [6]

Beck S., Barrel B.G.

Nature 331:269-272(1988).

818. (IGFBP) Insulin-like growth factor binding proteins signature

15 PROSITE cross-reference(s): PS00222; IGF_BINDING

The insulin-like growth factors (IGF-I and IGF-II) bind to specific binding proteins in extracellular fluids with high affinity [1,2,3]. These IGF-binding proteins (IGFBP) prolong the half-life of the IGFs and have been shown to either inhibit or stimulate the growth promoting effects of the IGFs on cells culture. They seem to alter the interaction of IGFs with their cell surface receptors. There are at least six different IGFBPs and they are structurally related.

25 The following growth-factor inducible proteins are structurally related to IGFBPs and could function as growth-factor binding proteins [4,5]:

- Mouse protein cyr61 and its probable chicken homolog, protein CEF-10.
- Human connective tissue growth factor (CTGF) and its mouse homolog, protein
- 30 FISP-12.
- Vertebrate protein NOV.

As a signature pattern a conserved cysteine-rich region located in the N-terminal

section of these proteins is used.

Consensus patternG-C-[GS]-C-C-x(2)-C-A-x(6)-C

Sequences known to belong to this class detected by the patternALL, except
5 for IGFBP-6's.

[1]

Rechler M.M.

Vitam. Horm. 47:1-114(1993).

[2]

Shimasaki S., Ling N.

Prog. Growth Factor Res. 3:243-266(1991).

[3]

Clemmons D.R.

Trends Endocrinol. Metab. 1:412-417(1990).

[4]

Bradham D.M., Igarashi A., Potter R.L., Grotendorst G.R.

J. Cell Biol. 114:1285-1294(1991).

[5]

Maloisel V., Martinerie C., Dambrine G., Plassiart G., Brisac M., Crochet

J., Perbal B.

Mol. Cell. Biol. 12:10-21(1992).

819. LMWPc : Low molecular weight phosphotyrosine protein phosphatase

25 Number of members: 34

[1]Medline: 94329182, The crystal structure of a low-molecular-weight phosphotyrosine
protein phosphatase. Su XD, Taddei N, Stefani M, Ramponi G, Nordlund P; Nature
1994;370:575-578.

820. (myosin_head) ATP/GTP-binding site motif A (P-loop)

PROSITE cross-reference(s): PS00017; ATP_GTP_A

From sequence comparisons and crystallographic data analysis it has been shown [1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP share a number of more or less conserved sequence motifs. The best conserved of these motifs is a glycine-rich region, which typically forms a flexible loop between a beta-strand and an alpha-helix. This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is generally referred to as the 'A' consensus sequence [1] or the 'P-loop' [5].

There are numerous ATP- or GTP-binding proteins in which the P-loop is found.

A number of protein families for which the relevance of the presence of such motif has been noted is listed below:

- ATP synthase alpha and beta subunits (see <PDOC00137>).
- Myosin heavy chains.
- Kinesin heavy chains and kinesin-like proteins (see <PDOC00343>).
- Dynamins and dynamin-like proteins (see <PDOC00362>).
- Guanylate kinase (see <PDOC00670>).
- Thymidine kinase (see <PDOC00524>).
- Thymidylate kinase (see <PDOC01034>).
- Shikimate kinase (see <PDOC00868>).
- Nitrogenase iron protein family (nifH/frxC) (see <PDOC00580>).
- ATP-binding proteins involved in 'active transport' (ABC transporters) [7] (see <PDOC00185>).
- DNA and RNA helicases [8,9,10].
- GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc.).
- Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.).
- Nuclear protein ran (see <PDOC00859>).
- ADP-ribosylation factors family (see <PDOC00781>).
- Bacterial dnaA protein (see <PDOC00771>).
- Bacterial recA protein (see <PDOC00131>).
- Bacterial recF protein (see <PDOC00539>).
- Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc.).
- DNA mismatch repair proteins mutS family (See <PDOC00388>).

- Bacterial type II secretion system protein E (see <PDOC00567>).

Not all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this is the case for tubulins or protein kinases. A special mention must be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.

Consensus pattern[AG]-x(4)-G-K-[ST]

[1]

Walker J.E., Saraste M., Runswick M.J., Gay N.J.
EMBO J. 1:945-951(1982).

[2]

Moller W., Amons R.
FEBS Lett. 186:1-7(1985).

[3]

Fry D.C., Kuby S.A., Mildvan A.S.
Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).

[4]

Dever T.E., Glynias M.J., Merrick W.C.
Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).

[5]

Saraste M., Sibbald P.R., Wittinghofer A.
Trends Biochem. Sci. 15:430-434(1990).

[6]

Koonin E.V.
J. Mol. Biol. 229:1165-1174(1993).

[7]

Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher

M.P.

J. Bioenerg. Biomembr. 22:571-592(1990).

[8]

Hodgman T.C.

5 Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

[9]

Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K.,
Schnier J., Slonimski P.P.

Nature 337:121-122(1989).

10 [10]

Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M.

Nucleic Acids Res. 17:4713-4730(1989).

821. PE: PE family

15 This family named after a PE motif near to the amino terminus of the domain. The PE family
of proteins all contain an amino-terminal region of about 110 amino acids. The carboxyl
terminus of this family are variable and fall into several classes. The largest class of PE
proteins is the highly repetitive PGRS class which have a high glycine content. The function
of these proteins is uncertain but it has been suggested that they may be related to antigenic
20 variation of Mycobacterium tuberculosis [1]. Number of members: 88

[1] Medline: 98295987. Deciphering the biology of Mycobacterium tuberculosis from the
complete genome sequence. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D,
Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D,
25 Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd
S, Hornsby T, Jagels K, Barrell BG, et al; Nature 1998;393:537-544.

822. (RNB) Ribonuclease II family signature

PROSITE cross-reference(s): PS01175; RIBONUCLEASE_II

30

On the basis of sequence similarities, the following bacterial and eukaryotic
proteins seem to form a family:

- *Escherichia coli* and related bacteria ribonuclease II (EC 3.1.13.1) (RNase II) (gene *rnb*) [1]. RNase II is an exonuclease involved in mRNA decay. It degrades mRNA by hydrolyzing single-stranded polyribonucleotides processively in the 3' to 5' direction.

5 - Bacterial protein *vacB*. In *Shigella flexneri*, *vacB* has been shown to be required for the expression of virulence genes at the posttranscriptional level.

- Yeast protein SSD1 (or SRK1) which is implicated in the control of the cell cycle G1 phase.

10 - Yeast protein DIS3 [2], which binds to ran (GSP1) and enhances the the nucleotide-releasing activity of RCC1 on ran.

- Fission yeast protein *dis3*, which is implicated in mitotic control.

- *Neurospora crassa* *cyt-4*, a mitochondrial protein required for RNA 5' and 3' end processing and splicing.

15 - Yeast protein MSU1, which is involved in mitochondrial biogenesis.

- *Synechocystis* strain PCC 6803 protein *zam* [3], which control resistance to the carbonic anhydrase inhibitor acetazolamide.

- *Caenorhabditis elegans* hypothetical protein F48E8.6.

20 The size of these proteins range from 644 residues (*rnb*) to 1250 (SSD1). While their sequence is highly divergent they share a conserved domain in their C-terminal section [4]. It is possible that this domain plays a role in a putative exonuclease function that would be common to all these proteins. A signature pattern was developed based on the core of this conserved domain.

25 Consensus pattern[HI]-[FYE]-[GSTAM]-[LIVM]-x(4,5)-Y-[STAL]-x-[FWVAC]-[TV]-[SA]-P-[LIVMA]-[RQ]-[KR]-[FY]-x-D-x(3)-[HQ]

[1]

30 Zilhao R., Camelo L., Arraiano C.M.
Mol. Microbiol. 8:43-51(1993).

[2]

Noguchi E., Hayashi N., Azuma Y., Seki T., Nakamura M., Nakashima N.,

Yanagida M., He X., Mueller U., Sazer S., Nishimoto T.
EMBO J. 15:5595-5605(1996).

[3]

Beuf L., Bedu S., Cami B., Joset F.

5 Plant Mol. Biol. 27:779-788(1995).

[4]

Mian I.S.

Nucleic Acids Res. 25:3187-3195(1997).

10 823. Src homology 2 (SH2) domain profile
PROSITE cross-reference(s): PS50001; SH2

The Src homology 2 (SH2) domain is a protein domain of about 100 amino-acid residues first identified as a conserved sequence region between the oncoproteins Src and Fps [1]. Similar sequences were later found in many other intracellular signal-transducing proteins [2]. SH2 domains function as regulatory modules of intracellular signalling cascades by interacting with high affinity to phosphotyrosine-containing target peptides in a sequence-specific and strictly phosphorylation-dependent manner [3,4,5,6].

20 The SH2 domain has a conserved 3D structure consisting of two alpha helices and six to seven beta-strands. The core of the domain is formed by a continuous beta-meander composed of two connected beta-sheets [7].

25 So far, SH2 domains have been identified in the following proteins:

- Many vertebrate, invertebrate and retroviral cytoplasmic (non-receptor) protein tyrosine kinases. In particular in the Src, Abl, Bkt, Csk and ZAP70 families of kinases.

30 - Mammalian phosphatidylinositol-specific phospholipase C gamma-1 and -2. Two copies of the SH2 domain are found in those proteins in between the catalytic 'X-' and 'Y-boxes' (see <PDOC50007>).

- Mammalian phosphatidyl inositol 3-kinase regulatory p85 subunit.

- Some vertebrate and invertebrate protein-tyrosine phosphatases.
- Mammalian Ras GTPase-activating protein (GAP).
- Adaptor proteins mediating binding of guanine nucleotide exchange factors to growth factor receptors: vertebrate GRB2, *Caenorhabditis elegans* sem-5 and *Drosophila* DRK.
- Mammalian Vav oncoprotein, a guanine-nucleotide exchange factor of the CDC24 family.
- Miscellaneous proteins interacting with vertebrate receptor protein tyrosine kinases: oncoprotein Crk, mammalian cytoplasmic proteins Nck, Shc.
- STAT proteins (signal transducers and activators of transcription).
- Chicken tensin.
- Yeast transcriptional control protein SPT6.

The profile developed to detect SH2 domains is based on a structural alignment consisting of 8 gap-free blocks and 7 linker regions totaling 92 match positions.

[1]

Sadowski I., Stone J.C., Pawson T.
Mol. Cell. Biol. 6:4396-4408(1986).

[2]

Russel R.B., Breed J., Barton G.J.
FEBS Lett. 304:15-20(1992).

[3]

Marangere L.E.M., Pawson T.
J. Cell Sci. Suppl. 18:97-104(1994).

[4]

Pawson T., Schlessinger J.
Curr. Biol. 3:434-442(1993).

[5]

Mayer B.J., Baltimore D.
Trends Cell. Biol. 3:8-13(1993).

[6]

Pawson T.

Nature 373:573-580(1995).

[7]

Kuriyan J., Cowburn D.

5 Curr. Opin. Struct. Biol. 3:828-837(1993).

824. Sulfate transporters signature

PROSITE cross-reference(s): PS01130; SULFATE_TRANSP

10 A number of proteins involved in the transport of sulfate across a membrane as well as some yet uncharacterized proteins have been shown [1,2] to be evolutionary related. These proteins are:

- 15
- Neurospora crassa sulfate permease II (gene cys-14).
 - Yeast sulfate permeases (genes SUL1 and SUL2).
 - Rat sulfate anion transporter 1 (SAT-1).
 - Mammalian DTDST, a probable sulfate transporter which, in Human, is involved in the genetic disease, diastrophic dysplasia (DTD).
 - Sulfate transporters 1, 2 and 3 from the legume Stylosanthes hamata.

20

 - Human pendrin (gene PDS), which is involved in a number of hearing loss genetic diseases.
 - Human protein DRA (Down-Regulated in Adenoma).
 - Soybean early nodulin 70.

25

 - Escherichia coli hypothetical protein ychM.
 - Caenorhabditis elegans hypothetical protein F41D9.5.

As expected by their transport function, these proteins are highly hydrophobic and seem to contain about 12 transmembrane domains. The best conserved region
30 seems to be located in the second transmembrane region and is used as a signature pattern.

Consensus pattern[PAV]-x-Y-[GS]-L-Y-[STAG](2)-x(4)-[LIVFYA]-[LIVST]-[YI]-

x(3)-[GA]-[GST]-S-[KR]

[1]

Sandal N.N., Marcker K.A.

5 Trends Biochem. Sci. 19:19-19(1994).

[2]

Smith F.W., Hawkesford M.J., Prosser I.M., Clarkson D.T.

Mol. Gen. Genet. 247:709-715(1995).

10 825. TYA: TYA transposon protein

Ty are yeast transposons. A 5.7kb transcript codes for p3 a fusion protein of TYA and TYB. The TYA protein is analogous to the gag protein of retroviruses. TYA a is cleaved to form 46kd protein which can form mature virion like particles [1]. Number of members: 59

15 [1] Medline: 97404699. Cryo-electron microscopy structure of yeast Ty retrotransposon virus-like particles. Palmer KJ, Tichelaar W, Myers N, Burns NR, Butcher SJ, Kingsman AJ, Fuller SD, Saibil HR; J Virol 1997;71:6863-6868.

826. Aldolase_II

20 Class II Aldolase and Adducin N-terminal domain.

-!- This family includes class II aldolases and adducins which have not been ascribed any enzymatic function. Number of members: 37

References:

25 [1] Medline: 93294819. The spatial structure of the class II L-fucose-1-phosphate aldolase from Escherichia coli. Dreyer MK, Schulz GE; J Mol Biol 1993;231:549-553.

[2] Medline: 96256522. Catalytic mechanism of the metal-dependent fucose aldolase from Escherichia coli as derived from the structure. Dreyer MK, Schulz GE; J Mol Biol 1996;259:458-466.

30

827. CBD_2

-!- Two tryptophan residues are involved in cellulose binding.

-!- Cellulose binding domain found in bacteria. Number of members: 51

References:

[1] Medline: 95284032. Solution structure of a cellulose-binding domain from *Cellulomonas fimi* by nuclear magnetic resonance spectroscopy. Xu GY, Ong E, Gilkes NR, Kilburn DG, Muhandiram DR, Harris-Brandts M, Carver JP, Kay LE, Harvey TS; Biochemistry 1995;34:6993-7009.

828. P

A unique feature of the eukaryotic subtilisin-like proprotein convertases is the presence of an additional highly conserved sequence of approximately 150 residues (P domain) located immediately downstream of the catalytic domain.

Number of members: 91

References:

[1] Medline: 94252314. A C-terminal domain conserved in precursor processing proteases is required for intramolecular N-terminal maturation of pro-Kex2 protease. Gluschankof P, Fuller RS; EMBO J 1994;13:2280-2288.

[2] Medline: 98225190. Regulatory roles of the P domain of the subtilisin-like prohormone convertases. Zhou A, Martin S, Lipkind G, LaMendola J, Steiner DF; J Biol Chem 1998;273:11107-11114.

829. Uncharacterized protein family UPF0020 signature

PROSITE cross-reference(s): PS01261; UPF0020

The following uncharacterized proteins have been shown [1] to share regions of similarities:

- *Escherichia coli* hypothetical protein ycbY and HI0116/15, the corresponding *Haemophilus influenzae* protein.
- *Bacillus subtilis* hypothetical protein ypsC.
- *Synechocystis* strain PCC 6803 hypothetical protein slr0064.
- *Methanococcus jannaschii* hypothetical proteins MJ0438 and MJ0710.

These are hydrophilic proteins of from 40 Kd to about 80 Kd. They can be

picked up in the database by the following pattern.

Consensus patternD-P-[LIVMF]-C-G-[ST]-G-x(3)-[LI]-E

5 References:

[1] Bairoch A. Unpublished observations (1997).

830. Uncharacterized protein family UPF0031 signatures

PROSITE cross-reference(s): PS01049; UPF0031_1; PS01050; UPF0031_2

10 The following uncharacterized proteins have been shown [1] to share regions of similarities:

- Yeast chromosome XI hypothetical protein YKL151c.
- Caenorhabditis elegans hypothetical protein R107.2.
- 15 - Escherichia coli hypothetical protein yjeF.
- Bacillus subtilis hypothetical protein yxkO.
- Helicobacter pylori hypothetical protein HP1363.
- Mycobacterium tuberculosis hypothetical protein MtCY77.05c.
- Mycobacterium leprae hypothetical protein B229_C2_201.
- 20 - Synechocystis strain PCC 6803 hypothetical protein sll1433.
- Methanococcus jannaschii hypothetical protein MJ1586.

These are proteins of about 30 to 40 Kd whose central region is well conserved. They can be picked up in the database by the following patterns.

25

Consensus pattern[SAV]-[IVW]-[LVA]-[LIV]-G-[PNS]-G-L-[GP]-x-[DENQT]

Consensus pattern[GA]-G-x-G-D-[TV]-[LT]-[STA]-G-x-[LIVM]

831. (ACOX)

30

Acyl-CoA oxidase

This is a family of Acyl-CoA oxidases EC:1.3.3.6. Acyl-coA oxidase converts acyl-CoA into trans-2-enoyl-CoA [1].

Number of members: 39

[1] Hayashi H, De Bellis L, Yamaguchi K, Kato A, Hayashi M, Nishimura M; Medline:
 5 98192624. "Molecular characterization of a glyoxysomal long chain acyl-CoA oxidase that is
 synthesized as a precursor of higher molecular mass in pumpkin." J Biol Chem
 1998;273:8301-8307.

10 832. (AICARFT_IMPCHas)
 AICARFT/IMPCHase bienzyme

This is a family of bifunctional enzymes catalysing the last steps in de novo purine
 biosynthesis. The bifunctional enzyme is found in both prokaryotes and eukaryotes. The
 15 second last step is catalysed by 5-aminoimidazole-4-carboxamide ribonucleotide
 formyltransferase EC:2.1.2.3 (AICARFT), this enzyme catalyses the formylation of AICAR
 with 10-formyl-tetrahydrofolate to yield FAICAR and tetrahydrofolate [1]. The last step is
 catalysed by IMP (Inosine monophosphate) cyclohydrolase EC:3.5.4.10 (IMPCHase),
 cyclizing FAICAR (5-formylaminoimidazole-4-carboxamide ribonucleotide) to IMP [1].

20 Number of members: 22

[1] Akira T, Komatsu M, Nango R, Tomooka A, Konaka K, Yamauchi M, Kitamura Y,
 Nomura S, Tsukamoto I; Medline: 97473523 "Molecular cloning and expression of a rat
 25 cDNA encoding 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP
 cyclohydrolase" [published erratum appears in Gene 1998 Feb 27;208(2):337] Gene
 1997;197:289-293.

[2] Rayl EA, Moroson BA, Beardsley GP; Medline: 96147205 "The human purH gene
 product, 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP
 30 cyclohydrolase. Cloning, sequencing, expression, purification, kinetic analysis, and domain
 mapping." J Biol Chem 1996;271:2225-2233.

833. (AOX)

Alternative oxidase

The alternative oxidase is used as a second terminal oxidase in the mitochondria, electrons are transferred directly from reduced ubiquinol to oxygen forming water [2]. This is not coupled to ATP synthesis and is not inhibited by cyanide, this pathway is a single step process [1]. In rice the transcript levels of the alternative oxidase are increased by low temperature [1].

Number of members: 27

[1] Ito Y, Saisho D, Nakazono M, Tsutsumi N, Hirai A; Medline: 98086211 "Transcript levels of tandem-arranged alternative oxidase genes in rice are increased by low temperature." Gene 1997;203:121-129.

[2] Li Q, Ritzel RG, McLean LL, McIntosh L, Ko T, Bertrand H, Nargang FE; Medline: 96366413 "Cloning and analysis of the alternative oxidase gene of *Neurospora crassa*." Genetics 1996;142:129-140.

834. (APH)

Protein kinases signatures and profile

Cross-reference(s): PS00107; PROTEIN_KINASE_ATP, PS00108; PROTEIN_KINASE_ST, PS00109; PROTEIN_KINASE_TYR, PS50011; PROTEIN_KINASE_DOM

Eukaryotic protein kinases [1 to 5] are enzymes that belong to a very extensive family of proteins which share a conserved catalytic core common to both serine/threonine and tyrosine protein kinases. There are a number of conserved regions in the catalytic domain of protein kinases. Two of these regions have been selected to build signature patterns. The first region, which is located in the N-terminal extremity of the catalytic domain, is a glycine-rich stretch of residues in the vicinity of a lysine residue, which has been shown to be involved in ATP

692

binding. The second region, which is located in the central part of the catalytic domain, contains a conserved aspartic acid residue which is important for the catalytic activity of the enzyme [6]; two signature patterns were derived for that region: one specific for serine/threonine kinases and the other for tyrosine kinases. A profile was developed which is based on the alignment in [1] and covers the entire catalytic domain.

Consensus pattern: [LIV]-G-{P}-G-{P}-[FYWMGSTNH]-[SGA]-{PW}-[LIVCAT]-{PD}-x-[GSTACLIVMFY]-x(5,18)-[LIVMFYWCSTAR]-[AIVP]-[LIVMFAGCKR]-K [K binds ATP]

Sequences known to belong to this class detected by the pattern the majority of known protein kinases but it fails to find a number of them, especially viral kinases which are quite divergent in this region and are completely missed by this pattern.

Consensus pattern: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT](3) [D is an active site residue]

Sequences known to belong to this class detected by the pattern. Most serine/ threonine specific protein kinases with 10 exceptions (half of them viral kinases) and also Epstein-Barr virus BGLF4 and Drosophila ninaC which have respectively Ser and Arg instead of the conserved Lys and which are therefore detected by the tyrosine kinase specific pattern described below.

Consensus pattern: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-[RSTAC]-x(2)-N-[LIVMFYC](3) [D is an active site residue] tyrosine specific protein kinases with the exception of human ERBB3 and mouse blk. This pattern will also detect most bacterial aminoglycoside phosphotransferases [8,9] and herpesviruses ganciclovir kinases [10]; which are proteins structurally and evolutionary related to protein kinases. Sequences known to belong to this class detected by the profile ALL, except for three viral kinases. This profile also detects receptor guanylate cyclases (see <PDOC00430>) and 2-5A-dependent ribonucleases. Sequence similarities between these two families and the eukaryotic protein kinase family have been noticed before. It also detects Arabidopsis thaliana kinase- like protein TMKL1 which seems to have lost its catalytic activity.

Note if a protein analyzed includes the two protein kinase signatures, the probability of it being a protein kinase is close to 100%. Note eukaryotic-type protein kinases have also been found in prokaryotes such as *Myxococcus xanthus* [11] and *Yersinia pseudotuberculosis*.

- 5 Note the patterns shown above has been updated since their publication in [7]. Note this documentation entry is linked to both signature patterns and a profile. As the profile is much more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so.

10 References

- [1] Hanks S.K., Hunter T., FASEB J. 9:576-596(1995).
[2] Hunter T., Meth. Enzymol. 200:3-37(1991).
[3] Hanks S.K., Quinn A.M., Meth. Enzymol. 200:38-62(1991).
[4] Hanks S.K., Curr. Opin. Struct. Biol. 1:369-383(1991).
5 [5] Hanks S.K., Quinn A.M., Hunter T., Science 241:42-52(1988).
[6] Knighton D.R., Zheng J., Ten Eyck L.F., Ashford V.A., Xuong N.-H., Taylor, S.S., Sowadski J.M., Science 253:407-414(1991).
[7] Bairoch A., Claverie J.-M., Nature 331:22(1988).
[8] Benner S., Nature 329:21-21(1987).
20 [9] Kirby R., J. Mol. Evol. 30:489-492(1992).
[10] Littler E., Stuart A.D., Chee M.S., Nature 358:160-162(1992).
[11] Munoz-Dorado J., Inouye S., Inouye M., Cell 67:995-1006(1991).

25 835. (Asp_Glu_race)

Aspartate and glutamate racemases signatures

Cross-reference(s) PS00923; ASP_GLU_RACEMASE_1 PS00924;
ASP_GLU_RACEMASE_2

30

Aspartate racemase (EC 5.1.1.13) and glutamate racemase (EC 5.1.1.3) are two evolutionary related bacterial enzymes that do not seem to require a cofactor for their activity [1].

Glutamate racemase, which interconverts L-glutamate into D-glutamate, is required for the

694

biosynthesis of peptidoglycan and some peptide-based antibiotics such as gramicidin S. In addition to characterized aspartate and glutamate racemases, this family also includes a hypothetical protein from *Erwinia carotovora* and one from *Escherichia coli* (ygeA). Two conserved cysteines are present in the sequence of these enzymes. They are expected to play a role in catalytic activity by acting as bases in proton abstraction from the substrate. Signature patterns were developed for both cysteines.

Consensus pattern: [IVA]-[LIVM]-x-C-x(0,1)-N-[ST]-[MSA]-[STH]-[LIVFYSTANK]

Consensus pattern: [LIVM](2)-x-[AG]-C-T-[DEH]-[LIVMFY]-[PNGRS]-x-[LIVM]

[1] Gallo K.A., Knowles J.R., Biochemistry 32:3981-3990(1993).

836. (ATP-sulfurylase)

ATP-sulfurylase

This family consists of ATP-sulfurylase or sulfate adenylyltransferase EC:2.7.7.4 some of which are part of a bifunctional polypeptide chain associated with adenosyl phosphosulphate (APS) kinase APS_kinase. Both enzymes are required for PAPS (phosphoadenosine-phosphosulfate) synthesis from inorganic sulphate [2]. ATP sulfurylase catalyses the synthesis of adenosine-phosphosulfate APS from ATP and inorganic sulphate [1].

Number of members: 37

[1] Kurima K, Warman ML, Krishnan S, Domowicz M, Krueger RC Jr, Deyrup A, Schwartz NB; Medline: 98337975 "A member of a family of sulfate-activating enzymes causes murine brachymorphism" [published erratum appears in Proc Natl Acad Sci U S A 1998 Sep 29;95(20):12071] Proc Natl Acad Sci U S A 1998;95:8681-8685.

[2] Rosenthal E, Leustek T; Medline: 96096529 "A multifunctional *Urechis caupo* protein, PAPS synthetase, has both ATP sulfurylase and APS kinase activities." Gene 1995;165:243-248.

837. (ATP-synt_F)

ATP synthase (F/14-kDa) subunit

5

This family includes 14-kDa subunit from vATPases [1], which is in the peripheral catalytic part of the complex [2]. The family also includes archaebacterial ATP synthase subunit F [3].

Number of members: 23

10

[1] Guo Y, Kaiser K, Wieczorek H, Dow JA; Medline: 96269411 "The *Drosophila melanogaster* gene *vha14* encoding a 14-kDa F-subunit of the vacuolar ATPase." *Gene* 1996;172:239-243.

[2] Peng SB, Crider BP, Tsai SJ, Xie XS, Stone DK; Medline: 96216416 "Identification of a 14-kDa subunit associated with the catalytic sector of clathrin-coated vesicle H⁺-ATPase." *J Biol Chem* 1996;271:3324-3327.

[3] Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Medline: 96324968 "Subunit structure and organization of the genes of the A1A0 ATPase from the Archaeon *Methanosarcina mazei* Go1." *J Biol Chem* 1996;271:18843-18852.

20

838. (CBD_4)

Starch binding domain

25 Number of members: 48

839. (CbiX)

30

The function of CbiX is uncertain, however it is found in cobalamin biosynthesis operons and so may have a related function. Some CbiX proteins contain a striking histidine-rich region at their C-terminus, which suggests that it might be involved in metal chelation [1].

Number of members: 6

[1] Raux E, Lanois A, Warren MJ, Rambach A, Thermes C: Medline: 98416126 "Cobalamin (vitamin B12) biosynthesis: identification and characterization of a *Bacillus megaterium* cobI operon." *Biochem J* 1998;335:159-166.

840. (Complex1_51K)

10 Respiratory-chain NADH dehydrogenase 51 Kd subunit signatures Cross-reference(s)
PS00644; COMPLEX1_51K_1 PS00645; COMPLEX1_51K_2

15 Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 51 Kd (in mammals), which is the second largest subunit of complex I and is a component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind to NAD, FMN, and a 2Fe-2S cluster.

20 The 51 Kd subunit is highly similar to [3,4]:

- Subunit alpha of *Alcaligenes eutrophus* NAD-reducing hydrogenase (gene *hoxF*) which also binds to NAD, FMN, and a 2Fe-2S cluster.
- Subunit NQO1 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase.
- 25 - Subunit F of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoF*).

30 The 51 Kd subunit and the bacterial hydrogenase alpha subunit contains three regions of sequence similarities. The first one most probably corresponds to the NAD-binding site, the second to the FMN-binding site, and the third one, which contains three cysteines, to the iron-sulfur binding region. Signature patterns have been developed for the FMN-binding and for the 2Fe-2S binding regions.

Consensus pattern: G-[AM]-G-[AR]-Y-[LIVM]-C-G-[DE](2)-[STA](2)-[LIM](2)-[EN]-S

Consensus pattern: E-S-C-G-x-C-x-P-C-R-x-G [The three C's are putative 2Fe-2S ligands]

- [1] Ragan C.L., Curr. Top. Bioenerg. 15:1-36(1987).
- [2] Weiss H., Friedrich T., Hofhaus G., Preis D., Eur. J. Biochem. 197:563-576(1991).
- 5 [3] Fearnley I.M., Walker J.E. Biochim. Biophys. Acta 1140:105-134(1992).
- [4] Weidner U., Geier S., Plock A., Friedrich T., Leif H., Weiss H., J. Mol. Biol. 233:109-122(1993).

10 841. (DAP_epimerase)

Diaminopimelate epimerase signature

Cross-reference(s) PS01326: DAP_EPIMERASE

15 Diaminopimelate epimerase (EC 5.1.1.7) catalyzes the isomeriazation of L,L- to D,L-meso-diaminopimelate in the biosynthetic pathway leading from aspartate to lysine. This enzyme is a protein of about 30 Kd. Two conserved cysteines seem [1] to function as the acid and base in the catalytic mechanism. As a signature pattern, the region surrounding the first of these two active site cysteines were selected.

20 Consensus pattern: N-x-D-G-S-x(4)-C-G-N-[GA]-x-R [C is an active site residue] Sequences known to belong to this class detected by the pattern ALL, except for an Anabaena dapF which has a Ser instead of the active site Cys.

- [1] Cirilli M., Zheng R., Scapin G., Blanchard J.S., Biochemistry 37:16452-16458(1998).

25

842. (DNA_gyraseB_C)

DNA topoisomerase II signature

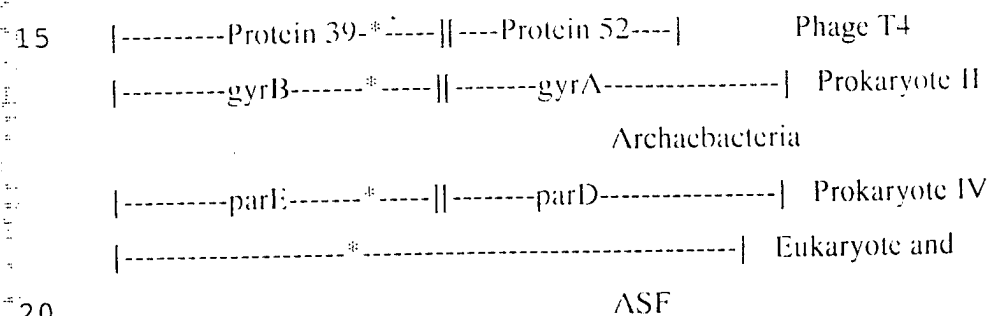
30 Cross-reference(s) PS00177; TOPOISOMERASE_II

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break.

Topoisomerase II is found in phages, archaebacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaebacteria the enzyme, known as DNA gyrase, consists of two subunits (genes *gyrA* and *gyrB* [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes *parC* and *parE*). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following representation:

<-----About-1400-residues----->



*: Position of the pattern.

As a signature pattern for this family of proteins, a region that contains a highly conserved pentapeptide was selected. The pattern is located in *gyrB*, in *parE*, and in protein 39 of phage T4 topoisomerase.

Consensus pattern: [LIVMA]-x-E-G-[DN]-S-A-x-[STAG]

- [1] Sternglanz R., Curr. Opin. Cell Biol. 1:533-535(1990).
- 30 [2] Bjornsti M.-A., Curr. Opin. Struct. Biol. 1:99-103(1991).
- [3] Sharma A., Mondragon A., Curr. Opin. Struct. Biol. 5:39-47(1995).
- [4] Roca J., Trends Biochem. Sci. 20:156-160(1995).

843. (DUF16)

Protein of unknown function

- 5 The function of this protein is unknown. It appears to only occur in *Mycoplasma pneumoniae*.

Number of members: 26

- 10 [1] Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R; Medline: 97105885
 "Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*."
 Nucleic Acids Res 1996;24:4420-4449.

15 844. (DUF21)

Domain of unknown function

20 This transmembrane region has no known function. Many of the sequences in this family are annotated as hemolysins, however this is due to a similarity to Swiss:Q54318 that does not contain this domain. This domain is found in the N-terminus of the proteins adjacent to two intracellular CBS domains CBS.

Number of members: 42

25

845. (DUF56)

Integral membrane protein

30

The members of this family are putative integral membrane proteins. The function of the family is unknown, however the family includes Sec59 from yeast. Sec59 is a dolichol

700

kinase EC:2.7.1.108, but it is not clear if the enzymatic activity resides in this region or its N terminal region.

Number of members: 13

5

846. (DUF94)

Domain of unknown function

10

The function of this domain is unknown. It is found in both eukaryotes and archaeobacteria. The alignment contains a completely conserved aspartate residue that may be functionally important. The eukaryotic domains contains three conserved cysteines and a histidine that might be metal binding, however these are absent in the archaeobacterial proteins.

5

Number of members: 9

847. (FF)

20

FF domain

This domain may be involved in protein-protein interaction [1].

25 Number of members: 42

[1] Bedford MT, Leder P; Medline: 99322199 "The FF domain: a novel motif that often accompanies WW domains." Trends Biochem Sci 1999;24:264-265.

30

848. (FLO_LFY)

Floricaula / Leafy protein

701

This family consists of various plant development proteins which are homologues of floricaula (FLO) and Leafy (LFY) proteins which are floral meristem identity proteins. Mutations in the sequences of these proteins affect flower and leaf development.

5 Number of members: 16

[1] Hofer J, Turner L, Hellens R, Ambrose M, Matthews P, Michael A, Ellis N; Medline: 97411151 "UNIFOLIATA regulates leaf and flower morphogenesis in pea." Curr Biol 1997;7:581-587.

10 [2] Weigel D, Alvarez J, Smyth DR, Yanofsky MF, Meyerowitz EM; Medline: 92274452 "LEAFY controls floral meristem identity in Arabidopsis." Cell 1992;69:843-859.

849. (G-patch)

15 G-patch domain

This domain is found in a number of RNA binding proteins, and is also found in proteins that contain RNA binding domains. This suggests that this domain may have an RNA binding function. This domain has seven highly conserved glycines.

20 Number of members: 47

[1] Aravind L, Koonin EV; Medline: 10470032 "G-patch: a new conserved domain in eukaryotic RNA-processing proteins and type D retroviral polyproteins." Trends Biochem Sci 1999;24:342-344.

850. (Gram-ve_porins)

General diffusion Gram-negative porins signature

30 Cross-reference(s) PS00576; GRAM_NEG_PORIN

The outer membrane of Gram-negative bacteria acts as a molecular filter for hydrophilic compounds. Proteins, known as porins [1], are responsible for the 'molecular sieve' properties

of the outer membrane. Porins form large water- filled channels which allows the diffusion of hydrophilic molecules into the periplasmic space. Some porins form general diffusion channels that allows any solutes up to a certain size (that size is known as the exclusion limit) to cross the membrane, while other porins are specific for a solute and contain a binding site for that solute inside the pores (these are known as selective porins). As porins are the major outer membrane proteins, they also serve as receptor sites for the binding of phages and bacteriocins. General diffusion porins generally assemble as trimer in the membrane and the transmembrane core of these proteins is composed exclusively of beta strands [2]. It has been shown [3] that a number of general porins are evolutionary related, these porins are:

- Enterobacteria phoE.
- Enterobacteria ompC.
- Enterobacteria ompF.
- Enterobacteria nmpC.
- Bacteriophage PA-2 LC.
- Neisseria PI.A.
- Neisseria PI.B.

As a signature pattern a conserved region was selected, located in the C-terminal part of these proteins, which spans two putative transmembrane beta strands.

Consensus pattern: [LIVMFY]-x(2)-G-x(2)-Y-x-F-x-K-x(2)-[SN]-[STAV]-[LIVMFYW]- V

[1] Benz R., Bauer K., Eur. J. Biochem. 176:1-19(1988).

[2] Jap B.K., Walian P.J., Q. Rev. Biophys. 23:367-403(1990).

[3] Jeanteur D., Lakey J.H., Pattus F., Mol. Microbiol. 5:2153-2164(1991).

851. (HlyD)

HlyD family secretion proteins signature

Cross-reference(s) PS00543; HLYD_FAMILY

Gram-negative bacteria produce a number of proteins which are secreted into the growth medium by a mechanism that does not require a cleaved N-terminal signal sequence. These

proteins, while having different functions, require the help of two or more proteins for their secretion across the cell envelope. Amongst which a protein belonging to the ABC transporters family (see the relevant entry <PDOC00185>) and a protein belonging to a family which is currently composed [1 to 5] of the following members:

5 Gene Species Protein which is exported

hlyD Escherichia coli Hemolysin

appD A.pleuropneumoniae Hemolysin

lcnD Lactococcus lactis Lactococcin A

10 lktD A.actinomycescomitans Leukotoxin

Pasteurella haemolytica

rtxD A.pleuropneumoniae Toxin-III

cyaD Bordetella pertussis Calmodulin-sensitive adenylate cyclase-
hemolysin (cyclolysin)

15 cvaA Escherichia coli Colicin V

prtE Erwinia chrysanthemi Extracellular proteases B and C

aprE Pseudomonas aeruginosa Alkaline protease

emrA Escherichia coli Drugs and toxins

yjcR Escherichia coli Unknown

20 These proteins are evolutionary related and consist of from 390 to 480 amino acid residues. They seem to be anchored in the inner membrane by a N-terminal transmembrane region. Their exact role in the secretion process is not yet known. The C-terminal section of these proteins is the best conserved region; a signature pattern from that region was derived.

25 Consensus pattern: [LIVM]-x(2)-G-[LM]-x(3)-[STGAV]-x-[LIVMT]-x-[LIVMT]-[GE]-x-[KR]-x-[LIVMFYW](2)-x-[LIVMFYW](3)

Sequences known to belong to this class detected by the pattern ALL, except for emrA and yjcR.

30 References:

[1] Gilson L., Mahanty H.K., Kolter R., EMBO J. 9:3875-3884(1990).

[2] Letoffe S., Delepelaire P., Wandersman C., EMBO J. 9:1375-1382(1990).

[3] Stoddard G.W., Petzel J.P., van Belkum M.J., Kok J., McKay L.L., Appl. Environ. Microbiol. 58:1952-1961(1992).

[4] Duong F., Lazdunski A., Cami B., Murgier M., Gene 121:47-54(1992).

[5] Lewis K., Trends Biochem. Sci. 19:119-123(1994).

5

852. (IBR)

In Between Ring fingers

10 The IBR (In Between Ring fingers) domain is found to occur between pairs of ring fingers (zf-C3HC4). The function of this domain is unknown. This domain has also been called the C6HC domain and DRIL (for double RING finger linked) domain [2].

Number of members: 25

15 [1] Morett E, Bork P; Medline: 10366851 "A novel transactivation domain in parkin."Trends Biochem Sci 1999;24:229-231.

[2] van der Reijden BA, Erpelinck-Verschueren CA, Lowenberg B, Jansen JH; Medline: 99349709 "TRIADS: a new class of proteins with a novel cysteine-rich signature." Protein Sci 1999;8:1557-1561.

853. (IPPT)

IPP transferase

25 [1] Durand JM, Bjork GR, Kuwae A, Yoshikawa M, Sasakawa C; Medline: 97440126 "The modified nucleoside 2-methylthio-N6-isopentenyladenosine in tRNA of Shigella flexneri is required for expression of virulence genes." J Bacteriol 1997;179:5777-5782.

[2] Boguta M, Hunter LA, Shen WC, Gillman EC, Martin NC, Hopper AK; Medline: 94187700 "Subcellular locations of MOD5 proteins: mapping of sequences sufficient for
30 targeting to mitochondria and demonstration that mitochondrial and nuclear isoforms commingle in the cytosol." Mol Cell Biol 1994;14:2298-2306.

[3] Gillman EC, Slusher LB, Martin NC, Hopper AK; Medline: 91203856 "MOD5 translation initiation sites determine N6-isopentenyladenosine modification of mitochondrial and cytoplasmic tRNA." Mol Cell Biol 1991;11:2382-2390.

5

854. (KE2)

KE2 family protein

10

The function of members of this family is unknown, although they have been suggested to contain a DNA binding leucine zipper motif [2].

Number of members: 9

15

[1] Ha H, Abe K, Artzt K; Medline: 92084131 "Primary structure of the embryo-expressed gene KE2 from the mouse H-2K region." Gene 1991;107:345-346.

[2] Shang HS, Wong SM, Tan HM, Wu M; Medline: 95129859 "YKE2, a yeast nuclear gene encoding a protein showing homology to mouse KE2 and containing a putative leucine-zipper motif." Gene 1994;151:197-201.

20

855. (Lipoprotein_6)

Prokaryotic membrane lipoprotein lipid attachment site

25

Cross-reference(s) PS00013; PROKAR_LIPOPROTEIN

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

30

- Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp).
- Escherichia coli lipoprotein-28 (gene nlpA).
- Escherichia coli lipoprotein-34 (gene nlpB).
- Escherichia coli lipoprotein nlpC.

- *Escherichia coli* lipoprotein nlpD.
- *Escherichia coli* osmotically inducible lipoprotein B (gene osmB).
- *Escherichia coli* osmotically inducible lipoprotein E (gene osmE).
- *Escherichia coli* peptidoglycan-associated lipoprotein (gene pal).
- 5 - *Escherichia coli* rare lipoproteins A and B (genes rplA and rplB).
- *Escherichia coli* copper homeostasis protein cutF (or nlpE).
- *Escherichia coli* plasmids traT proteins.
- *Escherichia coli* Col plasmids lysis proteins.
- A number of *Bacillus* beta-lactamases.
- 10 - *Bacillus subtilis* periplasmic oligopeptide-binding protein (gene oppA).
- *Borrelia burgdorferi* outer surface proteins A and B (genes ospA and ospB).
- *Borrelia hermsii* variable major protein 21 (gene vmp21) and 7 (gene vmp7).
- *Chlamydia trachomatis* outer membrane protein 3 (gene omp3).
- *Fibrobacter succinogenes* endoglucanase cel-3.
- 15 - *Haemophilus influenzae* proteins Pal and Pcp.
- *Klebsiella* pullulanase (gene pulA).
- *Klebsiella* pullulanase secretion protein pulS.
- *Mycoplasma hyorhinitis* protein p37.
- *Mycoplasma hyorhinitis* variant surface antigens A, B, and C (genes vlpABC).
- 20 - *Neisseria* outer membrane protein H.8.
- *Pseudomonas aeruginosa* lipopeptide (gene lppL).
- *Pseudomonas solanacearum* endoglucanase egl.
- *Rhodopseudomonas viridis* reaction center cytochrome subunit (gene cytC).
- *Rickettsia* 17 Kd antigen.
- 25 - *Shigella flexneri* invasion plasmid proteins mxiJ and mxiM.
- *Streptococcus pneumoniae* oligopeptide transport protein A (gene amiA).
- *Treponema pallidum* 34 Kd antigen.
- *Treponema pallidum* membrane protein A (gene tmpA).
- *Vibrio harveyi* chitinase (gene chb).
- 30 - *Yersinia* virulence plasmid protein yscJ.
- Halocyanin from *Natrobacterium pharaonis* [4], a membrane associated copper-binding protein. This is the first archaebacterial protein known to be modified in such a fashion).

From the precursor sequences of all these proteins, a consensus pattern and a set of rules to identify this type of post-translational modification were derived.

Consensus pattern: {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1)

The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence. Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be.

References

- [1] Hayashi S., Wu H.C., J. Bioenerg. Biomembr. 22:451-471(1990).
- [2] Klein P., Somorjai R.L., Lau P.C.K., Protein Eng. 2:15-20(1988).
- [3] von Heijne G., Protein Eng. 2:531-534(1989).
- [4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

856. (Lipoprotein_7)
Adhesin lipoprotein

This family consists of the p50 and variable adherence-associated antigen (Vaa) adhesins from *Mycoplasma hominis*. *M. hominis* is a mycoplasma associated with human urogenital diseases, pneumonia, and septic arthritis [1]. An adhesin is a cell surface molecule that mediates adhesion to other cells or to the surrounding surface or substrate. The Vaa antigen is a 50-kDa surface lipoprotein that has four tandem repetitive DNA sequences encoding a periodic peptide structure, and is highly immunogenic in the human host [1]. p50 is also a 50-kDa lipoprotein, having three repeats A,B and C, that may be a tetramer of 191-kDa in its native environment [2].

Number of members: 18

[1] Zhang Q, Wise KS; Medline: 96294788 "Molecular basis of size and antigenic variation of a *Mycoplasma hominis* adhesin encoded by divergent vaa genes. " Infect Immun 1996;64:2737-2744.

- 5 [2] Henrich B, Kitzerow A, Feldmann RC, Schaal H, Hadding U; Medline: 97047675 "Repetitive elements of the *Mycoplasma hominis* adhesin p50 can be differentiated by monoclonal antibodies." Infect Immun 1996;64:4027-4034.

10 857. (MaoC_like)
MaoC like domain

The MaoC protein is found to share similarity with a wide variety of enzymes; estradiol 17 beta-dehydrogenase 4, peroxisomal hydratase-dehydrogenase-epimerase, fatty acid synthase beta subunit. All these enzymes contain other domains. This domain is also present in the NodN nodulation protein N. No specific function has been assigned to this region of any of these proteins. The maoC gene is part of a operon with maoA which is involved in the synthesis of monoamine oxidase [1].

20 Number of members: 46

[1] Sugino H, Sasaki M, Azakami H, Yamashita M, Murooka Y Medline: 96235221 "A monoamine-regulated *Klebsiella aerogenes* operon containing the monoamine oxidase structural gene (maoA) and the maoC gene." J Bacteriol 1992;174:2485-2492.

25

858. (MSP)

Manganese-stabilizing protein / photosystem II polypeptide

- 30 This family consists of the 33 KDa photosystem II polypeptide from the oxygen evolving complex (OEC) of plants and cyanobacteria. The protein is also known as the manganese-stabilizing protein as it is associated with the manganese complex of the OEC and may provide the ligands for the complex [1].

Number of members: 17

[1] Philbrick JB, Zilinskas BA; Medline: 88334494 "Cloning, nucleotide sequence and mutational analysis of the gene encoding the Photosystem II manganese-stabilizing polypeptide of *Synechocystis* 6803." *Mol Gen Genet* 1988;212:418-425.

859. (NAC)

[1] Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV; Medline: 99342100 "Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell." *Genome Res* 1999;9:608-628.

Number of members: 27

860. (Nop)

Putative snoRNA binding domain

This family consists of various Pre RNA processing ribonucleoproteins. The function of the aligned region is unknown however it may be a common RNA or snoRNA or Nop1p binding domain. Nop5p (Nop58p) Swiss:Q12499 from yeast is the protein component of a ribonucleoprotein protein required for pre-18s rRNA processing and is suggested to function with Nop1p in a snoRNA complex [1]. Nop56p Swiss:O00567 and Nop5p interact with Nop1p and are required for ribosome biogenesis [2]. Prp31p Swiss:p49704 is required for pre-mRNA splicing in *S. cerevisiae* [3].

Number of members: 23

[1] Wu P, Brockenbrough JS, Metcalfe AC, Chen S, Aris JP; Medline: 98298165 "Nop5p is a small nucleolar ribonucleoprotein component required for pre- 18 S rRNA processing in yeast." J Biol Chem 1998;273:16453-16463.

[2] Gautier T, Berges T, Tollervy D, Hurt E; Medline: 8038777 "Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis." Mol Cell Biol 1997;17:7088-7098.

[3] Weidenhammer EM, Singh M, Ruiz-Noriega M, Woolford JL Jr; Medline: 96184869 "The PRP31 gene encodes a novel protein required for pre-mRNA splicing in *Saccharomyces cerevisiae*." Nucleic Acids Res 1996;24:1164-1170.

861. (Nramp)

Natural resistance-associated macrophage protein

The natural resistance-associated macrophage protein (NRAMP) family consists of Nramp1, Nramp2, and yeast proteins Smf1 and Smf2. The NRAMP family is a novel family of functional related proteins defined by a conserved hydrophobic core of ten transmembrane domains [5]. This family of membrane proteins are divalent cation transporters. Nramp1 is an integral membrane protein expressed exclusively in cells of the immune system and is recruited to the membrane of a phagosome upon phagocytosis [1]. By controlling divalent cation concentrations Nramp1 may regulate the interphagosomal replication of bacteria [1]. Mutations in Nramp1 may genetically predispose an individual to susceptibility to diseases including leprosy and tuberculosis conversely this might however provide protection from rheumatoid arthritis [1]. Nramp2 is a multiple divalent cation transporter for Fe²⁺, Mn²⁺ and Zn²⁺ amongst others it is expressed at high levels in the intestine; and is major transferrin-independent iron uptake system in mammals [1]. The yeast proteins Smf1 and Smf2 may also transport divalent cations [3].

Number of members: 36

[1] Govoni G, Gros P; Medline: 98383996 "Macrophage NRAMP1 and its role in resistance to microbial infections." Inflamm Res 1998;47:277-284.

711

[2] Agranoff DD, Krishna S Medline: 98294035 "Metal ion homeostasis and intracellular parasitism." Mol Microbiol 1998;28:403-412.

[3] Pinner E, Gruenheid S, Raymond M, Gros P; Medline: 98030569 "Functional complementation of the yeast divalent cation transporter family SMF by NRAMP2, a member of the mammalian natural resistance- associated macrophage protein family." J Biol Chem 1997;272:28933-28938.

[4] Cellier M, Belouchi A, Gros P; Medline: 96402487 "Resistance to intracellular infections: comparative genomic analysis of Nramp." Trends Genet 1996;12:201-204.

[5] Cellier M, Prive G, Belouchi A, Kwan T, Rodrigues V, Chia W, Gros P; Medline: 96036029 "Nramp defines a family of membrane proteins." Proc Natl Acad Sci U S A 1995;92:10089-10093.

862. (NTP_transf_2)

Nucleotidyltransferase domain

Members of this family belong to a large family of nucleotidyltransferases [1].

Number of members: 83

[1] Holm L, Sander C; Medline: 96005605 "DNA polymerase beta belongs to an ancient nucleotidyltransferase superfamily." Trends Biochem Sci 1995;20:345-347.

863. (Paramyxo_P)

Paramyxovirus P phosphoprotein

This family consists of paramyxovirus P phosphoprotein from sendai virus and human and bovine parainfluenza viruses. The P protein is an essential part of the viral RNA polymerase complex formed from the P and L proteins [1]. The exact role of the P protein in this complex is unknown but it is involved in multiple protein-protein interactions and binding the polymerase complex to the nucleocapsid or ribonucleoprotein template [1]. It also appears to

be important for the proper folding of the L protein [1]. The paramyxoviruses have a negative sense ssRNA genome [1].

Number of members: 15

5

[1] Bowman MC, Smallwood S, Moyer SA; Medline: 99329169 "Dissection of Individual Functions of the Sendai Virus Phosphoprotein in Transcription." J Virol 1999;73:6474-6483.

[2] Matsuoka Y, Curran J, Pelet T, Kolakofsky D, Ray R, Compans RW; Medline: 91237868 "The P gene of human parainfluenza virus type 1 encodes P and C proteins but not a cysteine-rich V protein." J Virol 1991;65:3406-3410.

10

864. (Patatin)

This family consists of various patatin glycoproteins from plants. The patatin protein accounts for up to 40% of the total soluble protein in potato tubers [2]. Patatin is a storage protein but it also has the enzymatic activity of lipid acyl hydrolase, catalysing the cleavage of fatty acids from membrane lipids [2].

Number of members: 21

[1] Banfalvi Z, Kostyal Z, Barta E; Medline: 95107249 "Solanum brevidens possesses a non-sucrose-inducible patatin gene." Mol Gen Genet 1994;245:517-522.

[2] Mignery GA, Pikaard CS, Park WD; Medline: 88226014 "Molecular characterization of the patatin multigene family of potato." Gene 1988;62:27-44.

865. (Pentapeptide_2)

Pentapeptide repeats (8 copies)

These repeats are found in many mycobacterial proteins. These repeats are most common in the PPE family of proteins, where they are found in the MPTR subfamily of PPE proteins.

The function of these repeats is unknown. The repeat can be approximately described as

30

713

XNXGX, where X can be any amino acid. These repeats are similar to Pentapeptide [1], however it is not clear if these two families are structurally related.

Number of members: 362

5

[1] Bateman A, Murzin A, Teichmann SA; Medline: 98318059 "Structure and distribution of pentapeptide repeats in bacteria." Protein Sci 1998;7:1477-1480.

[2] Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG; Medline: 98295987 "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence." Nature 1998;393:537-544.

10

15

866. (Peptidase_C13)

Peptidase C13 family

This family of peptidases is known as the hemoglobinase family because it contains a globin degrading enzyme from blood parasites Swiss:P42665. However relatives are found in plants and other organisms that have other functions. Members of this family are asparaginyl peptidases [1].

20

Number of members: 26

25

[1] Chen JM, Dando PM, Rawlings ND, Brown MA, Young NE, Stevens RA, Hewitt E, Watts C, Barrett AJ; Medline: 97218252 "Cloning, isolation, and characterization of mammalian legumain, an asparaginyl endopeptidase." J Biol Chem 1997;272:8090-8098.

30

867. (Pro_dh)

Proline dehydrogenase

Number of members: 25

[1] Ling M, Allen SW, Wood JM; Medline: 95055736 "Sequence analysis identifies the proline dehydrogenase and delta 1- pyrroline-5-carboxylate dehydrogenase domains of the multifunctional *Escherichia coli* PutA protein." J Mol Biol 1994;243:950-956.

868. (PsbP)

This family consists of the 23 kDa subunit of oxygen evolving system of photosystem II or PsbP from various plants (where it is encoded by the nuclear genome) and Cyanobacteria. The 23 KDa PsbP protein is required for PSII to be fully operational in vivo, it increases the affinity of the water oxidation site for Cl⁻ and provides the conditions required for high affinity binding of Ca²⁺ [2].

Number of members: 25

[1] Rova EM, Mc Ewen B, Fredriksson PO, Styring S; Medline: 97067138 "Photoactivation and photoinhibition are competing in a mutant of *Chlamydomonas reinhardtii* lacking the 23-kDa extrinsic subunit of photosystem II." J Biol Chem 1996;271:28918-28924.

[2] Kochhar A, Khurana JP, Tyagi AK; Medline: 97191538 "Nucleotide sequence of the psbP gene encoding precursor of 23-kDa polypeptide of oxygen-evolving complex in *Arabidopsis thaliana* and its expression in the wild-type and a constitutively photomorphogenic mutant." DNA Res 1996;3:277-285.

869. (PUA)

The PUA domain named after PseudoUridine synthase and Archaeosine transglycosylase, was detected in archaeal and eukaryotic pseudouridine synthases, archaeal archaeosine synthases, a family of predicted ATPases that may be involved in RNA modification, a family of predicted archaeal and bacterial rRNA methylases. Additionally, the PUA domain was detected in a family of eukaryotic proteins that also contain a domain homologous to the

translation initiation factor eIF1/SUI1; these proteins may comprise a novel type of translation factors. Unexpectedly, the PUA domain was detected also in bacterial and yeast glutamate kinases; this is compatible with the demonstrated role of these enzymes in the regulation of the expression of other genes [1]. It is predicted that the PUA domain is an RNA binding domain.

Number of members: 48

[1] Aravind L, Koonin EV; Medline: 99193178 "Novel predicted RNA-binding domains associated with the translation machinery." J Mol Evol 1999;48:291-302.

870. (RF1)

eRF1-like proteins

Members of this family are peptide chain release factors. The eukaryotic Release Factor 1 proteins (eRF1s) are involved in termination of translation. The eRF1 protein is functional for all stop codons and appears to abolish read-through of these codons. This family also includes other proteins for which the precise molecular function is unknown. Many of them are from Archaeobacteria. These proteins may also be involved in translation termination but this awaits experimental verification. Number of members: 25

[1] Frolova L, Le Goff X, Rasmussen HH, Cheperegin S, Dugeon G, Kress M, Arman I, Haenni AL, Celis JE, Philippe M, et al; Medline: 95082951 "A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor" [see comments] Nature 1994;372:701-703.

[2] Dugeon G, Jean-Jean O, Frolova L, Le Goff X, Philippe M, Kisselev L, Haenni AL; Medline: 97315314 "Eukaryotic release factor 1 (eRF1) abolishes readthrough and competes with suppressor tRNAs at all three termination codons in messenger RNA." Nucleic Acids Res 1997;25:2254-2258.

871. (Ribosomal_L14e)Ribosomal protein L14

716

This family includes the eukaryotic ribosomal protein L14.

Number of members: 15

- 5 872. (Ribosomal_S27)
Ribosomal protein S27a

This family of ribosomal proteins consists mainly of the 40S ribosomal protein S27a which is synthesized as a C-terminal extension of ubiquitin (CEP). The S27a domain compromises the C-terminal half of the protein. The synthesis of ribosomal proteins as extensions of ubiquitin promotes their incorporation into nascent ribosomes by a transient metabolic stabilization and is required for efficient ribosome biogenesis [3]. The ribosomal extension protein S27a contains a basic region that is proposed to form a zinc finger; its fusion gene is proposed as a mechanism to maintain a fixed ratio between ubiquitin necessary for degrading proteins and ribosomes a source of proteins [2].

Number of members: 36

- 20 873. (Spermine_synth)
Spermine/spermidine synthase

Spermine and spermidine are polyamines. This family includes spermidine synthase that catalyses the fifth (last) step in the biosynthesis of spermidine from arginine, and spermine synthase.

Number of members: 39

[1] Mezquita J, Pau M, Mezquita C; Medline: 97449308 "Characterization and expression of two chicken cDNAs encoding ubiquitin fused to ribosomal proteins of 52 and 80 amino acids." Gene 1997;195:313-319.

[2] Redman KL, Rechsteiner M; Medline: 89181932 "Identification of the long ubiquitin extension as ribosomal protein S27a." Nature 1989;338:438-440.

[3] Finley D, Bartel B, Varshavsky A; Medline: 89181925 "The tails of ubiquitin precursors are ribosomal proteins whose fusion to ubiquitin facilitates ribosome biogenesis." Nature 1989;338:394-401.

5

874. (Surp)

Surp module

10

[1] Denhez F, Lafyatis R; Medline: 94266805 "Conservation of regulated alternative splicing and identification of functional domains in vertebrate homologs to the Drosophila splicing regulator, suppressor-of-white-apricot." J Biol Chem 1994;269:16170-16179.

15

This domain is also known as the SWAP domain. SWAP stands for Suppressor-of-White-APricot. It has been suggested that these domains may be RNA binding [1].

Number of members: 32

20

875. (TFIIE)

TFIIE alpha subunit

25

The general transcription factor TFIIE has an essential role in eukaryotic transcription initiation together with RNA polymerase II and other general factors. Human TFIIE consists of two subunits TFIIE-alpha Swiss:P29083 and TFIIE-beta Swiss:P29084 and joins the preinitiation complex after RNA polymerase II and TFIIF [1]. This family consists of the conserved amino terminal region of eukaryotic TFIIE-alpha [2] and proteins from archaeobacteria that are presumed to be TFIIE-alpha subunits also Swiss:O29501 [3].

30

Number of members: 12

[1] Ohkuma Y, Sumimoto H, Hoffmann A, Shimasaki S, Horikoshi M, Roeder RG; Medline: 92065982 "Structural motifs and potential sigma homologies in the large subunit of human general transcription factor TFIIE." Nature 1991;354:398-401.

[2] Ohkuma Y, Hashimoto S, Roeder RG, Horikoshi M; Medline: 93087200 Identification of two large subdomains in TFIIE-alpha on the basis of homology between *Xenopus* and human sequences. *Nucleic Acids Res* 1992;20:5838-5838.

[3] Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Venter JC, et al; Medline: 98049343 "The complete genome sequence of the hyperthermophilic, sulphate- reducing archaeon *Archaeoglobus fulgidus*." *Nature* 1997;390:364-370.

876. (Transglut_core)

Cross-reference(s) PS00547; TRANSGLUTAMINASES

Transglutaminases (EC 2.3.2.13) (TGase) [1,2] are calcium-dependent enzymes that catalyze the cross-linking of proteins by promoting the formation of isopeptide bonds between the gamma-carboxyl group of a glutamine in one polypeptide chain and the epsilon-amino group of a lysine in a second polypeptide chain. TGases also catalyze the conjugation of polyamines to proteins. The best known transglutaminase is blood coagulation factor XIII, a plasma tetrameric protein composed of two catalytic A subunits and two non-catalytic B subunits. Factor XIII is responsible for cross-linking fibrin chains, thus stabilizing the fibrin clot. Other forms of transglutaminases are widely distributed in various organs, tissues and body fluids. Sequence data is available for the following forms of TGase:

- Transglutaminase K (Tgase K), a membrane-bound enzyme found in mammalian epidermis and important for the formation of the cornified cell envelope (gene TGM1).
- Tissue transglutaminase (TGase C), a monomeric ubiquitous enzyme located in the cytoplasm (gene TGM2).
- Transglutaminase 3, responsible for the later stages of cell envelope formation in the epidermis and the hair follicle (gene TGM3).
- Transglutaminase 4 (gene TGM4).

A conserved cysteine is known to be involved in the catalytic mechanism of TGases. The erythrocyte membrane band 4.2 protein, which probably plays an important role in regulating the shape of erythrocytes and their mechanical properties, is evolutionary related to TGases. However the active site cysteine is substituted by an alanine and the 4.2 protein does not show TGase activity.

Consensus pattern:[GT]-Q-[CA]-W-V-x-[SA]-[GA]-[IVT]-x(2)-T-x-[LMSC]-R-[CSA]-[LV]-G [The first C is the active site residue] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT NONE.

- [1] Ichinose A., Bottenus R.E., Davie E.W. J. Biol. Chem. 265:13411-13414(1990).
[2] Greenberg C.S., Birckbichler P.J., Rice R.H. FASEB J. 5:3071-3077(1991).

877. (TruB_N)

TruB family pseudouridylate synthase (N terminal domain)

Members of this family are involved in modifying bases in RNA molecules. They carry out the conversion of uracil bases to pseudouridine. This family includes TruB, a pseudouridylate synthase that specifically converts uracil 55 to pseudouridine in most tRNAs. This family also includes Cbf5p that modifies rRNA [2].

Number of members: 33

- [1] Nurse K, Wrzesinski J, Bakin A, Lane BG, Ofengand J; Medline: 96079944 "Purification, cloning, and properties of the tRNA psi 55 synthase from Escherichia coli." RNA 1995;1:102-112.
[2] Lafontaine DLJ, Bousquet-Antonelli C, Henry Y, Caizergues-Ferrer M, Tollervey D; Medline: 98139521 "The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase." Genes Dev 1998;12:527-537.

720

878. (UDPGP)

UTP--glucose-1-phosphate uridylyltransferase

This family consists of UTP--glucose-1-phosphate uridylyltransferases, EC:2.7.7.9. Also known as UDP-glucose pyrophosphorylase (UDPGP) and Glucose-1-phosphate uridylyltransferase. UTP--glucose-1-phosphate uridylyltransferase catalyses the interconversion of MgUTP + glucose-1-phosphate and UDP-glucose + MgPPi [1]. UDP-glucose is an important intermediate in mammalian carbohydrate interconversion involved in various metabolic roles depending on tissue type [1]. In Dictyostelium (slime mold) mutants in this enzyme abort the development cycle [2]. Also within the family is UDP-N-acetylglucosamine Swiss:Q16222 or AGX1 [3] and two hypothetical proteins from Borrelia burgdorferi the lyme disease spirochaete Swiss:O51893 and Swiss:O51036.

Number of members: 18

[1] Duggleby RG, Chao YC, Huang JG, Peng HL, Chang HY; Medline: 96202932 "Sequence differences between human muscle and liver cDNAs for UDPglucose pyrophosphorylase and kinetic properties of the recombinant enzymes expressed in Escherichia coli." Eur J Biochem 1996;235:173-179.

[2] Ragheb JA, Dottin RP; Medline: 87231075 "Structure and sequence of a UDP glucose pyrophosphorylase gene of Dictyostelium discoideum." Nucleic Acids Res 1987;15:3891-3906.

[3] Mio T, Yabe T, Arisawa M, Yamada-Okabe H; Medline: 98269105 "The eukaryotic UDP-N-acetylglucosamine pyrophosphorylases. Gene cloning, protein expression, and catalytic mechanism. J Biol Chem 1998;273:14392-14397.

879. (UPF004)

Uncharacterized protein family UPF0044 signature

Cross-reference(s) PS01301; UPF0044

The following uncharacterized proteins have been shown [1] to be highly similar:

- Bacillus subtilis hypothetical protein yqeI.

721

- Escherichia coli hypothetical protein yhbY and HI1333, the corresponding Haemophilus influenzae protein.

- Methanococcus jannaschii hypothetical protein MJ0652.

These are small proteins of 10 to 15 Kd. They can be picked up in the database

5 by the following pattern. This pattern is located in the N-terminal part of these proteins.

Consensus pattern: L-[ST]-x(3)-K-x(3)-[KR]-[SGA]-x-[GA]-H-x-L-x-P-[LIV]-x(2)- [LIV]-
[GA]-x(2)-G Sequences known to belong to this class detected by the patternALL. Other
10 sequence(s) detected in SWISS-PROT NONE.

880. (zf-A20)

A20-like zinc finger

15 A20- (an inhibitor of cell death)-like zinc fingers. The zinc finger mediates self-association in A20. These fingers also mediate IL-1-induced NF-kappa B activation.

Number of members: 22

20 [1] Heyninck K, Beyaert R; Medline: 99126071 "The cytokine-inducible zinc finger protein A20 inhibits IL-1-induced NF- kappaB activation at the level of TRAF6. FEBS Lett 1999;442:147-150.

[2] De Valck D, Heyninck K, Van Crielinge W, Contreras R, Beyaert R, Fiers W; Medline: 96390831 "A20, an inhibitor of cell death, self-associates by its
25 zinc finger domain." FEBS Lett 1996;384:61-64.

[3] Song HY, Rothe M, Goeddel DV; Medline: 96270609 "The tumor necrosis factor-inducible zinc finger protein A20 interacts with TRAF1/TRAF2 and inhibits NF-kappaB activation. Proc Natl Acad Sci U S A 1996;93:6721-6725.

30 [4] Opipari AW Jr, Boguski MS, Dixit VM; Medline: 90368626 "The A20 cDNA induced by tumor necrosis factor alpha encodes a novel type of zinc finger protein." J Biol Chem 1990;265:14705-14708.

881. (zf-PARP)

Poly(ADP-ribose) polymerase zinc finger domain

5 Cross-reference(s) PS00347; PARP_ZN_FINGER_1 PS50064; PARP_ZN_FINGER_2

10 Poly(ADP-ribose) polymerase (EC 2.4.2.30) (PARP) [1,2] is a eukaryotic enzyme that catalyzes the covalent attachment of ADP-ribose units from NAD(+) to various nuclear acceptor proteins. This post-translational modification of nuclear proteins is dependent on DNA. It appears to be involved in the regulation of various important cellular processes such as differentiation, proliferation and tumor transformation as well as in the regulation of the molecular events involved in the recovery of the cell from DNA damage. Structurally, PARP, about 1000 amino-acids residues long, consists of three distinct domains: an N-terminal zinc-dependent DNA-binding domain, a central automodification domain and a C-terminal NAD-binding domain. The DNA-binding region contains a pair of zinc finger domains which have been shown to bind DNA in a zinc-dependent manner. The zinc finger domains of PARP seem to bind specifically to single-stranded DNA. DNA ligase III [3] contains, in its N-terminal section, a single copy of a zinc finger highly similar to those of PARP.

20 Consensus pattern: C-[KR]-x-C-x(3)-I-x-K-x(3)-[RG]-x(16,18)-W-[FYH]-H-x(2)-C [The three C's and the H are zinc ligands] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE. Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-
25 PROTNONE.

Note: This documentation entry is linked to both signature patterns and a profile. As the profile is much more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so.

30 [1] Althaus F.R., Richter C.R. Mol. Biol. Biochem. Biophys. 37:1-126(1987).

[2] de Murcia G., Menissier de Murcia J. Trends Biochem. Sci. 19:172-176(1994).

[3] Wei Y.-F., Robins P., Carter K., Caldecott K., Pappin D.J.C., Yu G.-L., Wang R.-P., Shell B.K., Nash R.A., Schar P., Barnes D.E., Haseltine W.A., Lindahl T. Mol. Cell. Biol. 15:3206-3216(1995).

5 882. Adenylylsulfate kinase (APS_kinase)

Enzyme that catalyses the phosphorylation of adenylylsulfate to 3'-phosphoadenylylsulfate. This domain contains an ATP binding P-loop motif. Number of members: 34

10 [1] MacRae IJ, Rose AB, Segel IH; Medline: 99003196 "Adenosine 5'-phosphosulfate kinase from *Penicillium chrysogenum*. site- directed mutagenesis at putative phosphoryl-accepting and ATP P-loop residues. J Biol Chem 1998;273:28583-28589.

15 883. DNA polymerase family B signature DNA_POLYMERASE_B (DNA_pol_B)

Replicative DNA polymerases (EC 2.7.7.7) are the key enzymes catalyzing the accurate replication of DNA. They require either a small RNA molecule or a protein as a primer for the de novo synthesis of a DNA chain. On the basis of sequence similarity, a number of DNA polymerases have been grouped [1 to 7] under the designation of DNA polymerase family B. These are:

- 20
- Higher eukaryotes polymerases alpha.
 - Higher eukaryotes polymerases delta.
 - Yeast polymerase I/alpha (gene POL1), polymerase II/epsilon (gene POL2), polymerase III/delta (gene POL3) and polymerase REV3.
 - Escherichia coli polymerase II (gene dinA or polB).
 - 25 - Archaeobacterial polymerases.
 - Polymerases of viruses from the herpesviridae family.
 - Polymerases from Adenoviruses.
 - Polymerases from Baculoviruses.
 - Polymerases from Chlorella viruses.
 - 30 - Polymerases from Poxviruses.
 - Bacteriophage T4 polymerase.
 - Podoviridae bacteriophages Phi-29, M2 and PZA polymerase.
 - Tectiviridae bacteriophage PRD1 polymerase.

- Polymerases encoded on mitochondrial linear DNA plasmids in various fungi and plants (Kluyveromyces lactis pGKL1 and pGKL2, Agaricus bitorquis pEM, Ascobolus immersus pAI2, Claviceps purpurea pCLK1, Neurospora Kalilo and Maranhar, maize S-1, etc).

- 5 Six regions of similarity (numbered from I to VI) are found in all or a subset of the above polymerases. The most conserved region (I) includes a conserved tetrapeptide with two aspartate residues. Its function is not yet known. However, it has been suggested [3] that it may be involved in binding a magnesium ion. This conserved region was selected as a signature for this family of DNA polymerases.

10

Consensus pattern [YA]-[GLIVMSTAC]-D-T-D-[SG]-[LIVMFTC]-x-[LIVMSTAC]
Sequences known to belong to this class detected by the patternALL, except for yeast polymerase II/epsilon, Agaricus bitorquis pEM and Sulfolobus solfataricus polymerase II.

11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
217

725

- Terminal deoxynucleotidyltransferase (TdT) (EC 2.7.7.31). TdT catalyzes the elongation of polydeoxynucleotide chains by terminal addition. One of the functions of this enzyme is the addition of nucleotides at the junction of rearranged Ig heavy chain and T cell receptor gene segments during the maturation of B and T cells.

- 5 - African Swine Fever virus protein O174L [4].
- Fission yeast hypothetical protein SpAC2F7.06c.

These enzymes are small (about 40 Kd) compared with other polymerases and their reaction mechanism operates via a distributive mode, i.e. they dissociate from the template-primer after addition of each nucleotide.

As a signature pattern for this family of DNA polymerases, a highly conserved region that contains a conserved arginine and two conserved aspartic acid residues were selected. The latter together with the arginine have been shown [5] to be involved in primer binding in polymerase beta.

Consensus pattern G-[SG]-[LFY]-x-R-[GE]-x(3)-[SGCL]-x-D-[LIVM]-D- [LIVMFY](3)-x(2)-[SAP] Sequences known to belong to this class detected by the patternALL.

[1] Ito J., Braithwaite D.K. Nucleic Acids Res. 19:4045-4057(1991).

[2] Matsukage A., Nishikawa K., Ooi T., Seto Y., Yamaguchi M. J. Biol. Chem. 262:8960-8962(1987).

[3] Prasad R., Widen S.G., Singhal R.K., Watkins J., Prakash L., Wilson S.H. Nucleic Acids Res. 21:5301-5307(1993).

[4] Yanez R.J., Rodriguez J.M., Nogal M.L., Yuste L., Enriquez C., Rodriguez J.F., Vinuela E. Virology 208:249-278(1995).

[5] Date T., Yamamoto S., Tanihara K., Nishimoto Y., Matsukage A. Biochemistry 30:5286-5292(1991).

885. DUF14 - Domain of unknown function

This domain is found in glutamate synthase, tungsten formylmethanofuran dehydrogenase subunit c (FwdC) and molybdenum formylmethanofuran dehydrogenase subunit c (FmdC).

It has no known function. Number of members: 52

[1] Hochheimer A, Hedderich R, Thauer RK; Medline: 99035764. "The formylmethanofuran dehydrogenase isoenzymes in *Methanobacterium wolfei* and *Methanobacterium thermoautotrophicum*: induction of the molybdenum isoenzyme by molybdate and constitutive synthesis of the tungsten isoenzyme." Arch Microbiol 1998;170:389-393.

886. DUF18-Domain of unknown function

This domain of unknown function is found in several *C. elegans* proteins. The domain is 120 amino acids long and rich in cysteine residues. There are 16 conserved cysteine positions in the domain. Number of members: 34

887. DUF27-Domain of unknown function

This domain is found in a number of otherwise unrelated proteins. This domain is found at the C-terminus of the macro-H2A histone protein Swiss:Q02874. This domain is found in the non-structural proteins of several types of ssRNA viruses such as NSP2 from alphaviruses Swiss:P03317. This domain is also found on its own in a family of proteins from bacteria Swiss:P75918, archaeobacteria Swiss:O59182 and eukaryotes Swiss:Q17432, suggesting that it is involved in an important and ubiquitous cellular process. Number of members: 66

888. DUF37-Domain of unknown function

This domain is found in short (70 amino acid) hypothetical proteins from various bacteria. The domain contains three conserved cysteine residues. Swiss:Q44066 from *Aeromonas hydrophila* has been found to have hemolytic activity (unpublished). Number of members: 19

889. EGF-like domain signatures. (EGF-like)

A sequence of about thirty to forty amino-acid residues long found in the sequence of epidermal growth factor (EGF) has been shown [1 to 6] to be present, in a more or less conserved form, in a large number of other, mostly animal proteins. The proteins currently known to contain one or more copies of an EGF-like pattern are listed below.

- Adipocyte differentiation inhibitor (gene PREF-1) from mouse (6 copies).
- Agrin, a basal lamina protein that causes the aggregation of acetylcholine receptors on cultured muscle fibers (4 copies).

- Amphiregulin, a growth factor (1 copy).
- Betacellulin, a growth factor (1 copy).
- Blastula proteins BP10 and Span from sea urchin which are thought to be involved in pattern formation (1 copy).

- 5 - BM86, a glycoprotein antigen of cattle tick (7 copies).
- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity (1-2 copies). Homologous proteins are found in sea urchin - suBMP (1 copy) - and in Drosophila - the dorsal-ventral patterning protein tolloid (2 copies).
- 10 - Caenorhabditis elegans developmental proteins lin-12 (13 copies) and glp-1 (10 copies).
- Caenorhabditis elegans APX-1 protein, a patterning protein (4.5 copies).
- Calcium-dependent serine proteinase (CASP) which degrades the extracellular matrix proteins type I and IV collagen and fibronectin (1 copy).
- Cartilage matrix protein CMP (1 copy).
- 15 - Cartilage oligomeric matrix protein COMP (4 copies).
- Cell surface antigen 114/A10 (3 copies).
- Cell surface glycoprotein complex transmembrane subunit ASGP-2 from rat (2 copies).
- Coagulation associated proteins C, Z (2 copies) and S (4 copies).
- Coagulation factors VII, IX, X and XII (2 copies).
- 20 - Complement C1r components (1 copy).
- Complement C1s components (1 copy).
- Complement-activating component of Ra-reactive factor (RARF) (1 copy).
- Complement components C6, C7, C8 alpha and beta chains, and C9 (1 copy).
- Crumbs, an epithelial development protein from Drosophila (29 copies).
- 25 - Epidermal growth factor precursor (7-9 copies).
- Exogastrula-inducing peptides A, C, D and X from sea urchin (1 copy).
- Fat protein, a Drosophila cadherin-related tumor suppressor (5 copies).
- Fetal antigen 1, a probable neuroendocrine differentiation protein, which is derived from the delta-like protein (DLK) (6 copies).
- 30 - Fibrillin 1 (47 copies) and fibrillin 2 (14 copies).
- Fibropellins IA (21 copies), IB (13 copies), IC (8 copies), II (4 copies) and III (8 copies) from the apical lamina - a component of the extracellular matrix - of sea urchin.
- Fibulin-1 and -2, two extracellular matrix proteins (9-11 copies).

- Giant-lens protein (protein Argos), which regulates cell determination and axon guidance in the *Drosophila* eye (1 copy).

- Growth factor-related proteins from various poxviruses (1 copy).

- Gurken protein, a *Drosophila* developmental protein (1 copy).

5 - Heparin-binding EGF-like growth factor (HB-EGF), transforming growth factor alpha (TGF-alpha), growth factors Lin-3 and Spitz (1 copy); the precursors are membrane proteins, the mature form is located extracellular.

- Hepatocyte growth factor (HGF) activator (EC 3.4.21.-) (2 copies).

10 - LDL and VLDL receptors, which bind and transport low-density lipoproteins and very low-density lipoproteins (3 copies).

- LDL receptor-related protein (LRP), which may act as a receptor for endocytosis of extracellular ligands (22 copies).

- Leucocyte antigen CD97 (3 copies), cell surface glycoprotein EMR1 (6 copies) and cell surface glycoprotein F4/80 (7 copies).

15 - Limulus clotting factor C, which is involved in hemostasis and host defense mechanisms in Japanese horseshoe crab (1 copy).

- Meprin A alpha subunit, a mammalian membrane-bound endopeptidase (1 copy).

- Milk fat globule-EGF factor 8 (MFG-E8) from mouse (2 copies).

- Neuregulin GGF-I and GGF-II, two human glial growth factors (1 copy).

20 - Neurexins from mammals (3 copies).

- Neurogenic proteins Notch, Xotch and the human homolog Tan-1 (36 copies), Delta (9 copies) and the similar differentiation proteins Lag-2 from *Caenorhabditis elegans* (2 copies), Serrate (14 copies) and Slit (7 copies) from *Drosophila*.

- Nidogen (also called entactin), a basement membrane protein from chordates (2-6 copies).

25 - Ookinete surface proteins (24 Kd, 25 Kd, 28 Kd) from *Plasmodium* (4 copies).

- Pancreatic secretory granule membrane major glycoprotein GP2 (1 copy).

- Perforin, which lyses non-specifically a variety of target cells (1 copy).

- Proteoglycans aggrecan (1 copy), versican (2 copies), perlecan (at least 2 copies), brevican (1 copy) and chondroitin sulfate proteoglycan (gene PG-M) (2 copies).

30 - Prostaglandin G/H synthase 1 and 2 (EC 1.14.99.1) (1 copy), which is found in the endoplasmatic reticulum.

- S1-5, a human extracellular protein whose ultimate activity is probably modulated by the environment (5 copies).

- Schwannoma-derived growth factor (SDGF), an autocrine growth factor as well as a mitogen for different target cells (1 copy).

- Selectins. Cell adhesion proteins such as ELAM-1 (E-selectin), GMP-140 (P-selectin), or the lymph-node homing receptor (L-selectin) (1 copy).

5 - Serine/threonine-protein kinase homolog (gene Pro25) from *Arabidopsis thaliana*, which may be involved in assembly or regulation of light-harvesting chlorophyll A/B protein (2 copies).

- Sperm-egg fusion proteins PH-30 alpha and beta from guinea pig (1 copy).

- Stromal cell derived protein-1 (SCP-1) from mouse (6 copies).

10 - TDGF-1, human teratocarcinoma-derived growth factor 1 (1 copy).

- Tenascin (or neuronectin), an extracellular matrix protein from mammals (14.5 copies), chicken (TEN-A) (13.5 copies) and the related proteins human tenascin-X (18 copies) and tenascin-like proteins TEN-A and TEN-M from *Drosophila* (8 copies).

15 - Thrombomodulin (fetomodulin), which together with thrombin activates protein C (6 copies).

- Thrombospondin 1, 2 (3 copies), 3 and 4 (4 copies), adhesive glycoproteins that mediate cell-to-cell and cell-to-matrix interactions.

- Thyroid peroxidase 1 and 2 (EC 1.11.1.8) from human (1 copy).

- Transforming growth factor beta-1 binding protein (TGF-B1-BP) (16 or 18 copies).

20 - Tyrosine-protein kinase receptors Tek and Tie (EC 2.7.1.112) (3 copies).

- Urokinase-type plasminogen activator (EC 3.4.21.73) (UPA) and tissue plasminogen activator (EC 3.4.21.68) (TPA) (1 copy).

- Uromodulin (Tamm-horsfall urinary glycoprotein) (THP) (3 copies).

25 - Vitamin K-dependent anticoagulants protein C (2 copies) and protein S (4 copies) and the similar protein Z, a single-chain plasma glycoprotein of unknown function (2 copies).

- 63 Kd sperm flagellar membrane protein from sea urchin (3 copies).

- 93 Kd protein (gene nel) from chicken (5 copies).

- Hypothetical 337.6 Kd protein T20G5.3 from *Caenorhabditis elegans* (44 copies).

30 The functional significance of EGF domains in what appear to be unrelated proteins is not yet clear. However, a common feature is that these repeats are found in the extracellular domain of membrane-bound proteins or in proteins known to be secreted (exception: prostaglandin G/H synthase). The EGF domain includes six cysteine residues which have been shown (in

730

EGF) to be involved in disulfide bonds. The main structure is a two-stranded beta-sheet followed by a loop to a C-terminal short two-stranded sheet. Subdomains between the conserved cysteines strongly vary in length as shown in the following schematic representation of the EGF-like domain:

```

5      +-----+      +-----+      |      |      |
| x(4)-C-x(0,48)-C-x(3,12)-C-x(1,70)-C-x(1,6)-C-x(2)-G-a-x(0,21)-G-x(2)-C-x  |
|      *****
|      +-----+

```

10 'C': conserved cysteine involved in a disulfide bond.

'G': often conserved glycine

'a': often conserved aromatic amino acid

'*': position of both patterns.

'x': any residue

5 The region between the 5th and 6th cysteine contains two conserved glycines of which at least one is present in most EGF-like domains. Two patterns were created for this domain, each including one of these C-terminal conserved glycine residues.

20 Consensus pattern: C-x-C-x(5)-G-x(2)-C [The 3 C's are involved in disulfide bonds]

Sequences known to belong to this class detected by the pattern A majority, but not those that have very long or very short regions between the last 3 conserved cysteines of their EGF-like domain(s). Other sequence(s) detected in SWISS-PROT87 proteins, of which 27 can be considered as possible candidates.

25 Consensus pattern: C-x-C-x(2)-[GP]-[FYW]-x(4,8)-C [The three C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern A majority, but not those that have very long or very short regions between the last 3 conserved cysteines of their EGF-like domain(s). Other sequence(s) detected in SWISS-PROT83 proteins, of which 49
30 can be considered as possible candidates. Note The beta chain of the integrin family of proteins contains 2 cysteine- rich repeats which were said to be dissimilar with the EGF pattern [7].

731

Note Laminin EGF-like repeats (see <PDOC00961>) are longer than the average EGF module and contain a further disulfide bond C-terminal of the EGF-like region. Perlecan and agrin contain both EGF-like domains and laminin-type EGF-like domains. Note the pattern do not detect all of the repeats of proteins with multiple EGF-like repeats. Note see
5 <PDOC00913> for an entry describing specifically the subset of EGF-like domains that bind calcium.

[1] Davis C.G. New Biol. 2:410-419(1990).

[2] Blomquist M.C., Hunt L.T., Barker W.C. Proc. Natl. Acad. Sci. U.S.A. 81:7363-
10 7367(1984).

[3] Barker W.C., Johnson G.C., Hunt L.T., George D.G. Protein Nucl. Acid Enz. 29:54-
68(1986).

[4] Doolittle R.F., Feng D.F., Johnson M.S. Nature 307:558-560(1984).

[5] Appella E., Weber I.T., Blasi F. FEBS Lett. 231:1-4(1988).

[6] Campbell I.D., Bork P. Curr. Opin. Struct. Biol. 3:385-392(1993).

[7] Tamkun J.W., DeSimone D.W., Fonda D., Patel R.S., Buck C., Horwitz A.F., Hynes
R.O. Cell 46:271-282(1986).

890. Ham1 family (Ham1p_like)

This family consists of the HAM1 protein Swiss:P47119 and hypothetical archaeal bacterial and C. elegans proteins. HAM1 controls 6-N-hydroxylaminopurine (HAP) sensitivity and mutagenesis in S. cerevisiae Swiss:P47119 [1]. The HAM1 protein protects the cell from HAP, either on the level of deoxynucleoside triphosphate or the DNA level by a yet
25 unidentified set of reactions [1]. Number of members: 19

[1] Noskov VN, Staak K, Shcherbakova PV, Kozmin SG, Negishi K, Ono BC, Hayatsu H, Pavlov YI; Medline: 96381244 "HAM1, the gene controlling 6-N-hydroxylaminopurine sensitivity and mutagenesis in the yeast Saccharomyces cerevisiae." Yeast 1996;12:17-29.

891. (HCO3_cotransp)

Anion exchange is a cellular transport function which contributes to the regulation of cell pH and volume. Anion exchangers are a family of functionally related proteins that contributes to these properties by maintaining the intracellular level of the two principal anions: chloride and HCO_3^- . The best characterized anion exchanger is the band 3 protein [1], which is an erythrocyte anion exchange membrane glycoprotein. Band 3 is a protein of about 900 amino acids which consists of a cytoplasmic N-terminal domain of about 400 residues and an hydrophobic C-terminal section of about 500 residues that contains at least ten transmembrane regions. The cytoplasmic domain provides binding sites for cytoskeletal proteins, while the integral membrane domain is responsible for anion transport. Band 3 protein is specific to erythroid cells, at least two other proteins [2] structurally and functionally related to band 3, are found in nonerythroid tissues:

- AE2 (or B3 related protein; B3RP), a protein of 1200 residues, which seems to be present in a variety of cell types including lymphoid, kidney, and choroid plexus.
- AE3, a protein of 1200 residues, which is specific to neurons.

Structurally AE2 and AE3 are very similar to band 3, the main difference being an extension of some 300 residues of the N-terminal domain in AE2 and AE3.

Two signature patterns were developed for these proteins. The first pattern is based on a conserved stretch of sequence that contains four clustered positive charged residues and which is located at the C-terminal extremity of the cytoplasmic domain, just before the first transmembrane segment from the integral domain. The second pattern is based on the perfectly conserved sequence of the fifth transmembrane segment; this segment contains a lysine, which is the covalent binding site for the isothiocyanate group of DIDS, an inhibitor of anion exchange.

Consensus pattern F-G-G-[LIVM](2)-[KR]-D-[LIVM]-[RK]-R-R-Y Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [FI]-L-I-S-L-I-F-I-Y-E-T-F-x-K-L Sequences known to belong to this class detected by the pattern ALL.

[1] Jay D., Cantley L. Annu. Rev. Biochem. 55:511-538(1986).

[2] Reithmeier R.A.F. Curr. Opin. Struct. Biol. 3:515-523(1993).

892. ATP phosphoribosyltransferase signature (HisG)

ATP phosphoribosyltransferase (EC 2.4.2.17) is the enzyme that catalyzes the first step in the biosynthesis of histidine in bacteria, fungi and plants. It is a protein of about 23 to 32 Kd. As a signature pattern a region located in the C-terminal part of this enzyme was selected.

Consensus pattern E-x(5)-G-x-[SAG]-x(2)-[IV]-x-D-[LIV]-x(2)-[ST]-G-x-T-[LM]

Sequences known to belong to this class detected by the pattern ALL.

893. HNH endonuclease (HNH)

Number of members: 56

[1] Shub DA, Goodrich-Blair H, Eddy SR; Medline: 95117127 "Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns." Trends Biochem Sci 1994;19:402-404.

[2] Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS; Medline: 98026854 "Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family." Nucleic Acids Res 1997;25:4626-4638.

[3] Gorbalenya AE; Medline: 95004046 "Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family." Protein Sci 1994;3:1117-1120.

894. NEUROHYPOPHYS_HORM (hormone5)

Oxytocin (or ocytocin) and vasopressin [1] are small (nine amino acid residues), structurally and functionally related neurohypophysial peptide hormones. Oxytocin causes contraction of the smooth muscle of the uterus and of the mammary gland while vasopressin has a direct antidiuretic action on the kidney and also causes vasoconstriction of the peripheral vessels.

Like the majority of active peptides, both hormones are synthesized as larger protein precursors that are enzymatically converted to their mature forms. Peptides belonging to this family are also found in birds, fish, reptiles and amphibians (mesotocin, isotocin, valitocin, glumitocin, aspartocin, vasotocin, seritocin, asvatocin, phasvatocin), in worms (annetocin), octopi (cephalotocin), locust (locupressin or neuropeptide F1/F2) and in molluscs

(conopressins G and S) [2]. The pattern developed to detect this category of peptides spans their entire sequence and includes four invariant amino acid residues.

Consensus pattern C-[LIFY](2)-x-N-[CS]-P-x-G [The two C's are linked by a disulfide bond]. Sequences known to belong to this class detected by the pattern ALL.

[1] Acher R., Chauvet J. Biochimie 70:1197-1207(1988).

[2] Chauvet J., Michel G., Ouedraogo Y., Chou J., Chait B.T., Acher R. Int. J. Pept. Protein Res. 45:482-487(1995).

895. 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK)

All organisms require reduced folate cofactors for the synthesis of a variety of metabolites. Most microorganisms must synthesize folate de novo because they lack the active transport system of higher vertebrate cells which allows these organisms to use dietary folates.

Enzymes involved in folate biosynthesis are therefore targets for a variety of antimicrobial agents such as trimethoprim or sulfonamides. 7,8-dihydro-6-hydroxymethylpterin-

pyrophosphokinase (EC 2.7.6.3) (HPPK) catalyzes the attachment of pyrophosphate to 6-hydroxymethyl-7,8-dihydropterin to form 6-hydroxymethyl-7,8-dihydropteridine

pyrophosphate. This is the first step in a three-step pathway leading to 7,8-dihydrofolate.

Bacterial HPPK (gene folK or sulD) [1] is a protein of 160 to 270 amino acids. In the lower eukaryote *Pneumocystis carinii*, HPPK is the central domain of a multifunctional folate synthesis enzyme (gene fas) [2]. As a signature for HPPK, a conserved region located in the central section of these enzymes was selected.

Consensus pattern [KRHD]-x-[GA]-[PSAE]-R-x(2)-D-[LIV]-D-[LIVM](2) Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.

[1] Talarico T.L., Ray P.H., Dev I.K., Merrill B.M., Dallas W.S. J. Bacteriol. 174:5971-5977(1992).

[2] Volpes F., Dyer M., Scaife J.G., Darby G., Stammers D.K., Delves C.J. Gene 112:213-218(1992).

896. Metalloenzyme superfamily (Metalloenzyme)

This family includes phosphopentomutase Swiss:P07651 and 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, Swiss:P37689. This family is also related to alk_phosphatase [1]. The alignment contains the most conserved residues that are probably involved in metal binding and catalysis. Number of members: 34

[1] Galperin MY, Bairoch A, Koonin EV; Medline: 99180418 "A superfamily of metalloenzymes unifies phosphopentomutase and cofactor-independent phosphoglycerate mutase with alkaline phosphatases and sulfatases." Protein Sci 1998;7:1829-1835.

897. Penicillin amidase (Penicil_amidase)

Penicillin amidase or penicillin acylase EC:3.5.1.11 catalyses the hydrolysis of benzylpenicillin to phenylacetic acid and 6-aminopenicillanic acid (6-APA) a key intermediate in the the synthesis of penicillins [1]. Also in the family is cephalosporin acylase Swiss:P07662 and Swiss:P29958 aculeacin A acylase which are involved in the synthesis of related peptide antibiotics. Number of members: 13

[1] Verhaert RM, Riemens AM, van der Laan JM, van Duin J, Quax WJ; Medline: 97438505 "Molecular cloning and analysis of the gene encoding the thermostable penicillin G acylase from *Alcaligenes faecalis*. Appl Environ Microbiol 1997;63:3412-3418.

[2] Duggleby HJ, Tolley SP, Hill CP, Dodson EJ, Dodson G, Moody PC; Medline: 95115804 "Penicillin acylase has a single-amino-acid catalytic centre." Nature 1995;373:264-268.

898. Phosphoribosyl-AMP cyclohydrolase (PRA-CH)

This enzyme catalyses the third step in the histidine biosynthetic pathway. It requires Zn ions for activity. Number of members: 13

[1] D'Ordine RL, Klem TJ, Davisson VJ; Medline: 99129952 "N1-(5'-phosphoribosyl)adenosine-5'-monophosphate cyclohydrolase: purification and characterization of a unique metalloenzyme. *Biochemistry* 1999;38:1537-1546.

5

899. Phosphoribosyl-ATP pyrophosphohydrolase (PRA-PH)

This enzyme catalyses the second step in the histidine biosynthetic pathway. Number of members: 32

10

[1] Keesey JK Jr, Bigelis R, Fink GR; Medline: 79216449 "The product of the *his4* gene cluster in *Saccharomyces cerevisiae*. A trifunctional polypeptide." *J Biol Chem* 1979 Aug 10;254:7427-7433.

[2] Bruni CB, Carlomagno MS, Formisano S, Paoletta G; Medline: 86310274 "Primary and secondary structural homologies between the *HIS4* gene product of *Saccharomyces cerevisiae* and the *hisIE* and *hisD* gene products of *Escherichia coli* and *Salmonella typhimurium*." *Mol Gen Genet* 1986;203:389-396.

900. Prokaryotic membrane lipoprotein lipid attachment site (PstS)

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

25

- Major outer membrane lipoprotein (murein-lipoproteins) (gene *lpp*).
- *Escherichia coli* lipoprotein-28 (gene *nlpA*).
- *Escherichia coli* lipoprotein-34 (gene *nlpB*).
- *Escherichia coli* lipoprotein *nlpC*.
- *Escherichia coli* lipoprotein *nlpD*.

30

- *Escherichia coli* osmotically inducible lipoprotein B (gene *osmB*).
- *Escherichia coli* osmotically inducible lipoprotein E (gene *osmE*).
- *Escherichia coli* peptidoglycan-associated lipoprotein (gene *pal*).
- *Escherichia coli* rare lipoproteins A and B (genes *rplA* and *rplB*).

737

- Escherichia coli copper homeostasis protein cutF (or nlpE).
- Escherichia coli plasmids traT proteins.
- Escherichia coli Col plasmids lysis proteins.
- A number of Bacillus beta-lactamases.

- 5
- Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA).
 - Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB).
 - Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7).
 - Chlamydia trachomatis outer membrane protein 3 (gene omp3).
 - Fibrobacter succinogenes endoglucanase cel-3.

- 10
- Haemophilus influenzae proteins Pal and Pcp.
 - Klebsiella pullulunase (gene pulA).
 - Klebsiella pullulunase secretion protein pulS.
 - Mycoplasma hyorhinis protein p37.
 - Mycoplasma hyorhinis variant surface antigens A, B, and C (genes vlpABC).
- 15
- Neisseria outer membrane protein H.8.
 - Pseudomonas aeruginosa lipopeptide (gene lppL).
 - Pseudomonas solanacearum endoglucanase egl.
 - Rhodopseudomonas viridis reaction center cytochrome subunit (gene cytC).
 - Rickettsia 17 Kd antigen.
- 20
- Shigella flexneri invasion plasmid proteins mxjJ and mxjM.
 - Streptococcus pneumoniae oligopeptide transport protein A (gene amiA).
 - Treponema pallidum 34 Kd antigen.
 - Treponema pallidum membrane protein A (gene tmpA).
 - Vibrio harveyi chitobiase (gene chb).

- 25
- Yersinia virulence plasmid protein yscJ.
 - Halocyanin from Natrobacterium pharaonis [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion).

From the precursor sequences of all these proteins, a consensus pattern was derived and a set of rules to identify this type of post-translational modification.

30

Consensus pattern {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first

seven positions of the sequence. Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be.

- 5 [1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).
 [2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).
 [3] von Heijne G. Protein Eng. 2:531-534(1989).
 [4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

10

901. Ribosome recycling factor (RRF)

The ribosome recycling factor (RRF / ribosome release factor) dissociates the ribosome from the mRNA after termination of translation, and is essential bacterial growth [1]. Thus
 15 ribosomes are "recycled" and ready for another round of protein synthesis. Number of members: 27

[1] Janosi L, Shimizu I, Kaji A; Medline: 94240115 "Ribosome recycling factor (ribosome releasing factor) is essential for bacterial growth." Proc Natl Acad Sci U S A 1994;91:4249-4253.

902. S-layer homology(SLH)

S-layers are paracrystalline mono-layered assemblies of (glyco)proteins which coat the
 25 surface of bacteria [1]. Several S-layer proteins and some other cell wall proteins contain one or more copies of a domain of about 50-60 residues, which has been called SLH (for S-layer homology) [2]. There is strong evidence that this domain serves as an anchor to the peptidoglycan [3]. The SLH domain has been found in:

- S-layer glycoprotein of *Acetogenium kivui* (3 copies).
- 30 - S-layer 125 Kd protein of *Bacillus sphaericus* (3 copies).
- S-layer protein of *Bacillus anthracis* (3 copies).
- S-layer protein of *Bacillus licheniformis* (3 copies).
- S-layer protein (HWP) from *Bacillus brevis* strain HPD31 (3 copies).

- Middle cell wall protein (MWP) from *Bacillus brevis* strain 47 (3 copies).
 - S-layer protein (p100) of *Thermus thermophilus* (1 copy).
 - Outer membrane protein Omp-alpha from *Thermotoga maritima* (1 copy).
 - Cellulosome anchoring protein (gene *ancA*), outer layer protein B (OlpB) and a further potential cell surface glycoprotein from *Clostridium thermocellum* (3 copies; the first copy is missing its N-terminal third which is appended to the end of the third copy; may have arisen by circular permutation).
 - Amylopullulanase (gene *amyB*) from *Thermoanaerobacter thermosulfurogenes* (3 copies)
 - Amylopullulanase (gene *aapT*) from *Bacillus* strain XAL-601 (3 copies).
 - Endoglucanase from *Bacillus* strain KSM-635 (3 copies).
 - Exoglucanase (gene *xynX*) from *Clostridium thermocellum* (3 copies).
 - Xylanase A (gene *xynA*) from *Thermoanaerobacter saccharolyticum* (2 copies; 3 copies if a frameshift is taken into account).
 - Protein involved in butirosin production (ButB) from *Bacillus circulans* (2 incomplete copies; 3 copies if three frameshifts are taken into account).
 - Two hypothetical proteins from *Synechocystis* strain PCC 6803 (1 copy each).
 - A hypothetical protein with sequence similarity to amylopullulanases found 3' of amylase gene from *Bacillus circulans* (fragment of 1 copy; 3 copies if two frameshifts are taken into account).
- SLH domains are found at the N- or C-termini of mature proteins. They occur in single copy followed by a predicted coiled coil domain, or in three contiguous copies. Structurally, the SLH domain is predicted to contain two alpha-helices flanking a beta strand. The SLH sequences are fairly divergent with an average identity of about 25%. It is however possible to build a sequence pattern that starts at the second position of the domain and that spans 3/4 of its length.

Consensus pattern[LVFYT]-x-[DA]-x(2,5)-[DN GSATPHY]-[FYWPDA]-x(4)-[LIV]-x(2)-[GTALV]-x(4,6)-[LIVFYC]-x(2)-G-x-[PGSTA]-x(2,3)-[MFYA]-x- [PGAV]-x(3,10)-[LIVMA]-[STKR]-[RY]-x-[EQ]-x-[STALIVM] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.

[1] Beveridge T.J. Curr. Opin. Struct. Biol. 4:204-212(1994).

[2] Lupas A., Engelhardt H., Peters J., Santarius U., Volker S., Baumeister W. J. Bacteriol. 176:1224-1233(1994).

[3] Lemaire M., Ohayon H., Gounon P., Fujino T., Beguin P. J. Bacteriol. 177:2451-2459(1995).

5

903. Queuine tRNA-ribosyltransferase (TGT)

This is a family of queuine tRNA-ribosyltransferases EC:2.4.2.29, also known as tRNA-guanine transglycosylase and guanine insertion enzyme. Queuine tRNA-ribosyltransferase modifies tRNAs for asparagine, aspartic acid, histidine and tyrosine with queuine. It catalyses the exchange of guanine-34 at the wobble position with 7-aminomethyl-7-deazaguanine, and the addition of a cyclopentenediol moiety to 7-aminomethyl-7-deazaguanine-34 tRNA; giving a hypermodified base queuine in the wobble position [1,2]. The aligned region contains a zinc binding motif C-x-C-x2-C-x29-H, and important tRNA and 7-aminomethyl-7deazaguanine binding residues [1]. Number of members: 27

[1] Romier C, Reuter K, Suck D, Ficner R; Medline: 96256303 "Crystal structure of tRNA-guanine transglycosylase: RNA modification by base exchange." EMBO J 1996;15:2850-2857.

[2] Garcia GA, Koch KA, Chong S; Medline: 93287116 "tRNA-guanine transglycosylase from Escherichia coli. Overexpression, purification and quaternary structure." J Mol Biol 1993;231:489-497.

904. ThiC Family (ThiC)

ThiC is found within the thiamine biosynthesis operon. ThiC is involved in pyrimidine biosynthesis [2]. ThiC catalyzes the substitution of the pyrophosphate of 2-methyl-4-amino-5-hydroxymethylpyrimidine pyrophosphate by 4-methyl-5-(beta-hydroxyethyl)thiazole phosphate to yield thiamine phosphate [3]. Number of members: 12

[1] Vander Horn PB, Backstrom AD, Stewart V, Begley TP; Medline: 93163063 "Structural genes for thiamine biosynthetic enzymes (thiCEFGH) in Escherichia coli K-12." J Bacteriol 1993;175:982-992.

[2] Begley TP, Downs DM, Ealick SE, McLafferty FW, Van Loon AP, Taylor S, Campobasso N, Chiu HJ, Kinsland C, Reddick JJ, Xi J; Medline: 99311269 "Thiamin biosynthesis in prokaryotes." Arch Microbiol 1999;171:293-300.

[3] Zhang Y, Taylor SV, Chiu HJ, Begley TP; Medline: 97284509 "Characterization of the Bacillus subtilis thiC operon involved in thiamine biosynthesis." J Bacteriol 1997;179:3030-3035.

905. Putative tRNA binding domain (tRNA_bind)

This domain is found in prokaryotic methionyl-tRNA synthetases, prokaryotic phenylalanyl tRNA synthetases the yeast GU4 nucleic-binding protein (G4p1 or p42, ARC1) [2], human tyrosyl-tRNA synthetase [1], and endothelial-monocyte activating polypeptide II. G4p1 binds specifically to tRNA form a complex with methionyl-tRNA synthetases [2]. In human tyrosyl-tRNA synthetase this domain may direct tRNA to the active site of the enzyme [2].

This domain may perform a common function in tRNA aminoacylation [1]. Number of members: 12

[1] Kleeman TA, Wei D, Simpson KL, First EA; Medline: 97306356 "Human tyrosyl-tRNA synthetase shares amino acid sequence homology with a putative cytokine." J Biol Chem 1997;272:14420-14425.

[2] Simos G, Segref A, Fasiolo F, Hellmuth K, Shevchenko A, Mann M, Hurt EC; Medline: 97050848 "The yeast protein Arc1p binds to tRNA and functions as a cofactor for the methionyl-and glutamyl-tRNA synthetases." EMBO J 1996;15:5437-5448.

906. UbiA prenyltransferase family signature (UbiA)

The following prenyltransferases are evolutionary related [1,2]:

- Bacterial 4-hydroxybenzoate octaprenyltransferase (gene ubiA).
- Yeast mitochondrial para-hydroxybenzoate--polyprenyltransferase (gene COQ2).
- Protoheme IX farnesyltransferase (heme O synthase) from yeast and mammals (gene COX10) and from bacteria (genes cyoE or ctaB).

These proteins probably contain seven transmembrane segments. The best conserved region is located in a loop between the second and third of these segments and was used as a signature pattern.

- 5 Consensus pattern N-x(3)-[DE]-x(2)-[LIF]-D-x(2)-[VM]-x-R-[ST]-x(2)-R-x(4)-G Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROTNONE.

[1] Melzer M., Heide L. Biochim. Biophys. Acta 1212:93-102(1994).

10 [2] Mogi T., Saiki K., Anraku Y. Mol. Microbiol. 14:391-398(1994).

907. Uncharacterized protein family UPF0044 signature (UPF0044)

The following uncharacterized proteins have been shown [1] to be highly similar:

- 15 - *Bacillus subtilis* hypothetical protein yqeI.
- *Escherichia coli* hypothetical protein yhbY and HI1333, the corresponding *Haemophilus influenzae* protein.
- *Methanococcus jannaschii* hypothetical protein MJ0652.

These are small proteins of 10 to 15 Kd. They can be picked up in the database by the following pattern. This pattern is located in the N-terminal part of these proteins.

20 Consensus pattern L-[ST]-x(3)-K-x(3)-[KR]-[SGA]-x-[GA]-H-x-L-x-P-[LIV]-x(2)-[LIV]-[GA]-x(2)-G Sequences known to belong to this class detected by the patternALL.

25 908. ATP synthase (C/AC39) subunit (vATP-synt_AC39)

This family includes the AC39 subunit from vacuolar ATP synthase Swiss:P32366 [1], and the C subunit from archaeobacterial ATP synthase [2]. The family also includes subunit C from the Sodium transporting ATP synthase from *Enterococcus hirae* Swiss:P43456 [3].

30 Number of members: 12

[1] Bauerle C, Ho MN, Lindorfer MA, Stevens TH; Medline: 93286119 "The *Saccharomyces cerevisiae* VMA6 gene encodes the 36-kDa subunit of the vacuolar H(+)-ATPase membrane sector." J Biol Chem 1993;268:12749-12757.

[2] Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Medline: 96324968

5 "Subunit structure and organization of the genes of the A1A0 ATPase from the Archaeon *Methanosarcina mazei* Go1." J Biol Chem 1996;271:18843-18852.

[3] Takase K, Kakinuma S, Yamato I, Konishi K, Igarashi K, Kakinuma Y; Medline: 94209269 "Sequencing and characterization of the ntp gene cluster for vacuolar- type Na(+)-translocating ATPase of *Enterococcus hirae*." J Biol Chem 1994;269:11037-11044.

10

909. ATP synthase (E/31 kDa) subunit (vATP-synt_E)

This family includes the vacuolar ATP synthase E subunit [1], as well as the archaebacterial ATP synthase E subunit [2]. Number of members: 24

[1] Foury F; Medline: 91009356 "The 31-kDa polypeptide is an essential subunit of the vacuolar ATPase in *Saccharomyces cerevisiae*." J Biol Chem 1990;265:18554-18560.

[2] Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Medline: 96324968
 15 "Subunit structure and organization of the genes of the A1A0 ATPase from the Archaeon *Methanosarcina mazei* Go1." J Biol Chem 1996;271:18843-18852.

910. (WW)

The WW domain [1-4,E1] (also known as rsp5 or WWP) has been originally discovered as a
 25 short conserved region in a number of unrelated proteins, among them dystrophin, the gene responsible for Duchenne muscular dystrophy. The domain, which spans about 35 residues, is repeated up to 4 times in some proteins. It has been shown [5] to bind proteins with particular proline- motifs, [AP]-P-P-[AP]-Y, and thus resembles somewhat SH3 domains. It appears to contain beta-strands grouped around four conserved aromatic positions; generally Trp. The
 30 name WW or WWP derives from the presence of these Trp as well as that of a conserved Pro. It is frequently associated with other domains typical for proteins in signal transduction processes.

Proteins containing the WW domain are listed below.

- Dystrophin, a multidomain cytoskeletal protein. Its longest alternatively spliced form consists of an N-terminal actin-binding domain, followed by 24 spectrin-like repeats, a cysteine-rich calcium-binding domain and a C-terminal globular domain. Dystrophin forms tetramers and is thought to have multiple functions including involvement in membrane stability, transduction of contractile forces to the extracellular environment and organization of membrane specialization. Mutations in the dystrophin gene lead to muscular dystrophy of Duchenne or Becker type. Dystrophin contains one WW domain C-terminal of the spectrin-repeats.

- Utrophin, a dystrophin-like protein of unknown function.

- Vertebrate YAP protein is a substrate of an unknown serine kinase. It binds to the SH3 domain of the Yes oncoprotein via a proline-rich region. This protein appears in alternatively spliced isoforms, containing either one or two WW domains [6].

- Mouse NEDD-4 plays a role in the embryonic development and differentiation of the central nervous system. It contains 3 WW modules followed by a HECT domain. The human ortholog contains 4 WW domains, but the third WW domain is probably spliced resulting in an alternate NEDD-4 protein with only 3 WW modules [3].

- Yeast RSP5 is similar to NEDD-4 in its molecular organization. It contains an N-terminal C2 domain (see <PDOC00380>, followed by a histidine-rich region, 3 WW domains and a HECT domain.

- Rat FE65, a transcription-factor activator expressed preferentially in liver. The activator domain is located within the N-terminal 232 residues of FE65, which also contain the WW domain.

- Yeast ESS1/PTF1, a putative peptidyl prolyl cis-trans isomerase from family ppiC (see <PDOC00840>). A related protein, dodo (gene dod) exists in Drosophila and in mammals (gene PIN1).

- Tobacco DB10 protein. The WW domain is located N-terminal to the region with similarity to ATP-dependent RNA helicases.

- IQGAP, a human GTPase activating protein acting on ras. It contains an N-terminal domain similar to fly muscle mp20 protein and a C-terminal ras GTPase activator domain.

- Yeast pre-mRNA processing protein PRP40, Caenorhabditis elegans ZK1098.1 and fission yeast SpAC13C5.02 are related proteins with similarity to MYO2-type myosin, each containing two WW-domains at the N-terminus.

745

- *Caenorhabditis elegans* hypothetical protein C38D4.5, which contains one WW module, a PH domain (see <PDOC50003>) and a C-terminal phosphatidylinositol 3-kinase domain.

- Yeast hypothetical protein YFL010c.

For the sensitive detection of WW domains, a profile was developed which spans the whole
5 homology region as well as a pattern.

Consensus pattern W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-x(2)-P

Sequences known to belong to this class detected by the pattern ALL. Other sequence(s)
detected in SWISS-PROT8. Sequences known to belong to this class detected by the
10 profile ALL.

[1] Bork P., Sudol M. Trends Biochem. Sci. 19:531-533(1994).

[2] Andre B., Springael J.Y. Biochem. Biophys. Res. Commun. 205:1201-1205(1994).

[3] Hofmann K.O., Bucher P. FEBS Lett. 358:153-157(1995).

[4] Sudol M., Chen H.I., Bougeret C., Einbond A., Bork P. FEBS Lett. 369:67-71(1995).

[5] Chen H.I., Sudol M. Proc. Natl. Acad. Sci. U.S.A. 92:7819-7823(1995).

[6] Sudol M., Bork P., Einbond A., Kastury K., Druck T., Negrini M., Huebner K., Lehman
D. J. Biol. Chem. 270:14733-14741(1995).

911. Xeroderma pigmentosum (XP) [1] (XPG_1)

Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a
high incidence of sunlight-induced skin cancer. People's skin cells with this condition are
hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair.

25 There are a minimum of seven genetic complementation groups involved in this pathway:
XP-A to XP-G. The defect in XP-G can be corrected by a 133 Kd nuclear protein called XPG
(or XPGC) [2].

XPG belongs to a family of proteins [2,3,4,5,6] that are composed of two main subsets:

30 - Subset 1, to which belongs XPG, RAD2 from budding yeast and rad13 from fission yeast.
RAD2 and XPG are single-stranded DNA endonucleases [7,8]. XPG makes the 3'incision in
human DNA nucleotide excision repair [9].

- Subset 2, to which belongs mouse and human FEN-1, rad2 from fission yeast, and RAD27 from budding yeast. FEN-1 is a structure-specific endonuclease.

In addition to the proteins listed in the above groups, this family also includes:

- 5 - Fission yeast exo1, a 5'→3' double-stranded DNA exonuclease that could act in a pathway that corrects mismatched base pairs.
- Yeast EXO1 (DHS1), a protein with probably the same function as exo1.
- Yeast DIN7.

10 Sequence alignment of this family of proteins reveals that similarities are largely confined to two regions. The first is located at the N-terminal extremity (N-region) and corresponds to the first 95 to 105 amino acids. The second region is internal (I-region) and found towards the C-terminus; it spans about 140 residues and contains a highly conserved core of 27 amino acids that includes a conserved pentapeptide (E-A-[DE]-A-[QS]). It is possible that the conserved acidic residues are involved in the catalytic mechanism of DNA excision repair in XPG. The amino acids linking the N- and I-regions are not conserved; indeed, they are largely absent from proteins belonging to the second subset.

20 Two signature patterns were developed for these proteins. The first corresponds to the central part of the N-region, the second to part of the I-region and includes the putative catalytic core pentapeptide.

25 Consensus pattern [VI]-[KRE]-P-x-[FYIL]-V-F-D-G-x(2)-[PIL]-x-[LVC]-K Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

30 Consensus pattern [GS]-[LIVM]-[PER]-[FYS]-[LIVM]-x-A-P-x-E-A-[DE]-[PAS]- [QS]-[CLM] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

[1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).

[2] Scherly D., Nospikel T., Corlet J., Ucla C., Bairoch A., Clarkson S.G. Nature 363:182-185(1993).

- [3] Carr A.M., Sheldrick K.S., Murray J.M., Al-Harithy R., Watts F.Z., Lehmann A.R. Nucleic Acids Res. 21:1345-1349(1993).
- [4] Murray J.M., Tavassoli M., Al-Harithy R., Sheldrick K.S., Lehmann A.R., Carr A.M., Watts F.Z. Mol. Cell. Biol. 14:4878-4888(1994).
- 5 [5] Harrington J.J., Lieber M.R. Genes Dev. 8:1344-1355(1994).
- [6] Szankasi P., Smith G.R. Science 267:1166-1169(1995).
- [7] Habraken Y., Sung P., Prakash L., Prakash S. Nature 366:365-368(1993).
- [8] O'Donovan A., Scherly D., Clarkson S.G., Wood R.D. J. Biol. Chem. 269:15965-15968(1994).
- 10 [9] O'Donovan A., Davies A.A., Moggs J.G., West S.C., Wood R.D. Nature 371:432-435(1994).

912. 5-formyltetrahydrofolate cyclo-ligase (5-FTHF_cyc-lig)

5-formyltetrahydrofolate cyclo-ligase or methenyl-THF synthetase EC:6.3.3.2 catalyses the interchange of 5-formyltetrahydrofolate (5-FTHF) to 5,10-methenyltetrahydrofolate, this requires ATP and Mg²⁺ [1]. 5-FTHF is used in chemotherapy where it is clinically known as Leucovorin [2].

Number of members: 23

[1] Dayan A, Bertrand R, Beauchemin M, Chahla D, Mamo A, Filion M, Skup D, Massie B, Jolivet J; Medline: 96096540 "Cloning and characterization of the human 5,10-methenyltetrahydrofolate synthetase-encoding cDNA." Gene 1995;165:307-311.

25 [2] Maras B, Stover P, Valiante S, Barra D, Schirch V; Medline: 94308074 "Primary structure and tetrahydropteroylglutamate binding site of rabbit liver cytosolic 5,10-methenyltetrahydrofolate synthetase." J Biol Chem 1994;269:18429-18433.

913. Cytosolic long-chain acyl-CoA thioester hydrolase (Acyl-CoA_hydro)

30 This family consist of various cytosolic long-chain acyl-CoA thioester hydrolases including human and rat [1,2]. The aligned region is repeated with in the sequence of human and rat cytosolic long-chain acyl-CoA thioester hydrolases of this family. Long-chain acyl-CoA

hydrolases hydrolyse palmitoyl-CoA to CoA and palmitate, they also catalyse the hydrolysis of other long chain fatty acyl-CoA thioesters. Long-chain acyl-CoA hydrolases are present in all living organisms and they may provide a mechanism for the control of lipid metabolism [1].

5 Number of members: 24

[1]Yamada J, Furihata T, Iida N, Watanabe T, Hosokawa M, Satoh T, Someya A, Nagaoka I, Suga T; Medline: 97236308 "Molecular cloning and expression of cDNAs encoding rat brain and liver cytosolic long-chain acyl-CoA hydrolases." Biochem Biophys Res Commun
10 1997;232:198-203.

[2] Broustas CG, Larkins LK, Uhler MD, Hajra AK; Medline: 96209964 "Molecular cloning and expression of cDNA encoding rat brain cytosolic acyl-coenzyme A thioester hydrolase." J Biol Chem 1996;271:10470-10476.

15 914. Agglutinin

Lectin (probable mannose binding)

Members of this family are plant lectins. Many if not all are mannose specific.

Number of members: 87

20 [1] Wright CS, Hester G; Medline: 97094989 "The 2.0 Å structure of a cross-linked complex between snowdrop lectin and a branched mannopentaose: evidence for two unique binding modes." Structure 1996;4:1339-1352.

25 915. (ANF_RECEPTORS)

Natriuretic peptides are hormones involved in the regulation of fluid and electrolyte homeostasis. These hormones stimulate the intracellular production of cyclic GMP as a second messenger.

30 Currently, three types of natriuretic peptide receptors are known [1,2]. Two express guanylate cyclase activity: GC-A (or ANP-A) which seems specific to atrial natriuretic peptide (ANP), and GC-B (or ANP-B) which seems to be stimulated more effectively by brain natriuretic

peptide (BNP) than by ANP. The third receptor (ANP-C) is probably responsible for the clearance of ANP from the circulation and does not play a role in signal transduction.

GC-A and GC-B are plasma membrane-bound proteins that share the following topology: an N-terminal extracellular domain which acts as the ligand binding region, then a transmembrane domain followed by a large cytoplasmic C- terminal region that can be subdivided into two domains: a protein kinase-like domain (see <PDOC00100>) that appears important for proper signalling and a guanylate cyclase catalytic domain (see <PDOC00425>). The topology of ANP-C is different: like GC-A and -B it possesses an extracellular ligand-binding region and a transmembrane domain, but its cytoplasmic domain is very short.

A pattern was developed from the ligand-binding region of natriuretic peptide receptors based on a highly conserved region located in the N-terminal part of the domain.

Consensus patternG-P-x-C-x-Y-x-A-A-x-V-x-R-x(3)-H-W Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT NONE.

[1] Garbers D.L. New Biol. 2:499-504(1990).

[2] Schulz S., Chinkers M., Garbers D.L. FASEB J. 2:2026-2035(1989).

916. (Apocytochrome)

Cytochrome c family heme-binding site signature

In proteins belonging to cytochrome c family [1], the heme group is covalently attached by thioether bonds to two conserved cysteine residues. The consensus sequence for this site is Cys-X-X-Cys-His and the histidine residue is one of the two axial ligands of the heme iron. This arrangement is shared by all proteins known to belong to cytochrome c family, which presently includes cytochromes c, c', c1 to c6, c550 to c556, cc3/Hmc, cytochrome f and reaction center cytochrome c.

750

Consensus pattern C-{CPWHF}-{CPWR}-C-H-{CFYW} Sequences known to belong to this class detected by the pattern ALL, except for four cytochrome c's which lack the first thioether bond. Other sequence(s) detected in SWISS-PROT 454.

- 5 Note: some cytochrome c's have more than a single bound heme group c4 has 2, c7 has 3, c3 has 4, the reaction center has 4, and cc3/Hmc has 16 !

[1] Mathews F.S. Prog. Biophys. Mol. Biol. 45:1-56(1985).

- 10 917. ATP-synt_A-c. ATP synthase Alpha chain, C terminal

[1] Medline: 94344236. Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria. Abrahams JP, Leslie AG, Lutter R, Walker JE; Nature 1994;370:621-628.

Number of members: 125

- 5 918. (Basic)

Myc-type, 'helix-loop-helix' dimerization domain signature

HELIX_LOOP_HELIX

A number of eukaryotic proteins, which probably are sequence specific DNA- binding proteins that act as transcription factors, share a conserved domain of 40 to 50 amino acid residues. It has been proposed [1] that this domain is formed of two amphipathic helices joined by a variable length linker region that could form a loop. This 'helix-loop-helix' (HLH) domain mediates protein dimerization and has been found in the proteins listed below

[2,3,E1,E2]. Most of these proteins have an extra basic region of about 15 amino acid residues that is adjacent to the HLH domain and specifically binds to DNA. They are referred as basic helix-loop-helix proteins (bHLH), and are classified in two groups: class A (ubiquitous) and class B (tissue-specific). Members of the bHLH family bind variations on the core sequence 'CANNTG', also referred to as the E-box motif. The homo- or

heterodimerization mediated by the HLH domain is independent of, but necessary for DNA

binding, as two basic regions are required for DNA binding activity. The HLH proteins

lacking the basic domain (Emc, Id) function as negative regulators since they form

heterodimers, but fail to bind DNA. The hairy-related proteins (hairy, E(spl), deadpan) also

repress transcription although they can bind DNA. The proteins of this subfamily act together with co-repressor proteins, like groucho, through their C-terminal motif WRPW.

- The myc family of cellular oncogenes [4], which is currently known to contain four members: c-myc [E3], N-myc, L-myc, and B-myc. The myc genes are thought to play a role in cellular differentiation and proliferation.

- Proteins involved in myogenesis (the induction of muscle cells). In mammals MyoD1 (Myf-3), myogenin (Myf-4), Myf-5, and Myf-6 (Mrf4 or herculin), in birds CMD1 (QMF-1), in *Xenopus* MyoD and MF25, in *Caenorhabditis elegans* CeMyoD, and in *Drosophila nautilus* (nau).

- Vertebrate proteins that bind specific DNA sequences ('E boxes') in various immunoglobulin chains enhancers: E2A or ITF-1 (E12/pan-2 and E47/pan-1), ITF-2 (tcf4), TFE3, and TFEB.

- Vertebrate neurogenic differentiation factor 1 that acts as differentiation factor during neurogenesis.

- Vertebrate MAX protein, a transcription regulator that forms a sequence-specific DNA-binding protein complex with myc or mad.

- Vertebrate Max Interacting Protein 1 (MXI1 protein) which acts as a transcriptional repressor and may antagonize myc transcriptional activity by competing for max.

- Proteins of the bHLH/PAS superfamily which are transcriptional activators. In mammals, AH receptor nuclear translocator (ARNT), single-minded homologs (SIM1 and SIM2), hypoxia-inducible factor 1 alpha (HIF1A), AH receptor (AHR), neuronal pas domain proteins (NPAS1 and NPAS2), endothelial pas domain protein 1 (EPAS1), mouse ARNT2, and human BMAL1. In *Drosophila*, single-minded (SIM), AH receptor nuclear translocator (ARNT), trachealess protein (TRH), and similar protein (SIMA).

- Mammalian transcription factors HES, which repress transcription by acting on two types of DNA sequences, the E box and the N box.

- Mammalian MAD protein (max dimerizer) which acts as transcriptional repressor and may antagonize myc transcriptional activity by competing for max.

- Mammalian Upstream Stimulatory Factor 1 and 2 (USF1 and USF2), which bind to a symmetrical DNA sequence that is found in a variety of viral and cellular promoters.

- Human lyl-1 protein; which is involved, by chromosomal translocation, in T-cell leukemia.

- Human transcription factor AP-4.

- Mouse helix-loop-helix proteins MATH-1 and MATH-2 which activate E box- dependent transcription in collaboration with E47.

- Mammalian stem cell protein (SCL) (also known as tal1), a protein which may play an important role in hemopoietic differentiation. SCL is involved, by chromosomal translocation, in stem-cell leukemia.

- Mammalian proteins Id1 to Id4 [5]. Id (inhibitor of DNA binding) proteins lack a basic DNA-binding domain but are able to form heterodimers with other HLH proteins, thereby inhibiting binding to DNA.

- Drosophila extra-macrochaetae (emc) protein, which participates in sensory organ patterning by antagonizing the neurogenic activity of the achaete- scute complex. Emc is the homolog of mammalian Id proteins.

- Human Sterol Regulatory Element Binding Protein 1 (SREBP-1), a transcriptional activator that binds to the sterol regulatory element 1 (SRE-1) found in the flanking region of the LDLR gene and in other genes.

- Drosophila achaete-scute (AS-C) complex proteins T3 (l'sc), T4 (scute), T5 (achaete) and T8 (asense). The AS-C proteins are involved in the determination of the neuronal precursors in the peripheral nervous system and the central nervous system.

- Mammalian homologs of achaete-scute proteins, the MASH-1 and MASH-2 proteins.

- Drosophila atonal protein (ato) which is involved in neurogenesis.

- Drosophila daughterless (da) protein, which is essential for neurogenesis and sex-determination.

- Drosophila deadpan (dpn), a hairy-like protein involved in the functional differentiation of neurons.

- Drosophila delilah (dei) protein, which is plays an important role in the differentiation of epidermal cells into muscle.

- Drosophila hairy (h) protein, a transcriptional repressor which regulates the embryonic segmentation and adult bristle patterning.

- Drosophila enhancer of split proteins E(spl), that are hairy-like proteins active during neurogenesis. also act as transcriptional repressors.

- Drosophila twist (twi) protein, which is involved in the establishment of germ layers in embryos.

- Maize anthocyanin regulatory proteins R-S and LC.

- Yeast centromere-binding protein 1 (CPF1 or CBF1). This protein is involved in chromosomal segregation. It binds to a highly conserved DNA sequence, found in centromeres and in several promoters.

- Yeast INO2 and INO4 proteins.

5 - Yeast phosphate system positive regulatory protein PHO4 which interacts with the upstream activating sequence of several acid phosphatase genes.

- Yeast serine-rich protein TYE7 that is required for ty-mediated ADH2 expression.

- *Neurospora crassa* nuc-1, a protein that activates the transcription of structural genes for phosphorus acquisition.

10 - Fission yeast protein esc1 which is involved in the sexual differentiation process.

The schematic representation of the helix-loop-helix domain is shown here:

XXXXXXXXXXXXXXXXXXXXXXXXXXXXX-----XXXXXXXXXXXXXXXXXXXXXXXXXXXXX Amphipathic
helix 1 Loop Amphipathic helix 2

The signature pattern that had been developed to detect this domain spans completely the second amphipathic helix.

Consensus pattern[DENSTAP]-[KR]-[LIVMAGSNT]-{FYWCPhKR}-[LIVMT]-[LIVM]-
x(2)-[STAV]-[LIVMSTACKR]-x-[VMFYH]-[LIVMTA]-{P}-{P}-[LIVMRKHQ]

Sequences known to belong to this class detected by the pattern the majority but far from all.
Other sequence(s) detected in SWISS-PROT135.

[1] Murre C., McCaw P.S., Baltimore D. Cell 56:777-783(1989).

25 [2] Garrel J., Campuzano S. BioEssays 13:493-498(1991).

[3] Kato G.J., Dang C.V. FASEB J. 6:3065-3072(1992).

[4] Krause M., Fire A., Harrison S.W., Priess J., Weintraub H. Cell 63:907-919(1990).

[5] Riechmann V., van Cruechten I., Sablitzky F. *Nucleic Acids Res.* 22:749-755(1994).

30 919. (Beta-lactamase)

Beta-lactamases classes -A, -C, and -D active site

Beta-lactamases (EC 3.5.2.6) [1,2] are enzymes which catalyze the hydrolysis of an amide bond in the beta-lactam ring of antibiotics belonging to the penicillin/cephalosporin family. Four kinds of beta-lactamase have been identified [3]. Class-B enzymes are zinc containing proteins whilst class -A, C and D enzymes are serine hydrolases. The three

5 classes of serine beta-lactamases are evolutionary related and belong to a superfamily [4] that also includes DD-peptidases and a variety of other penicillin-binding proteins (PBP's). All these proteins contain a Ser-x-x-Lys motif, where the serine is the active site residue. Although clearly homologous, the sequences of the three classes of serine beta-lactamases exhibit a large

10 degree of variability and only a small number of residues are conserved in addition to the catalytic serine.

Since a pattern detecting all serine beta-lactamases would also pick up many unrelated sequences, it was decided to provide specific patterns, centered on the active site serine, for each of the three classes.

Consensus pattern [FY]-x-[LIVMFY]-x-S-[TV]-x-K-x(4)-[AGLM]-x(2)-[LC] [S is the active site residue] Sequences known to belong to this class detected by the patternALL class-A beta-lactamases. Other sequence(s) detected in SWISS-PROT7.

Consensus pattern F-E-[LIVM]-G-S-[LIVMG]-[SA]-K [The first S is the active site residue] Sequences known to belong to this class detected by the patternALL class-C beta-lactamases. Other sequence(s) detected in SWISS-PROT NONE.

25 Consensus pattern [PA]-x-S-[ST]-F-K-[LIV]-[PAL]-x-[STA]-[LI] [S is the active site residue] Sequences known to belong to this class detected by the patternALL class-D beta-lactamases. Other sequence(s) detected in SWISS-PROT NONE.

[1] Ambler R.P. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 289:321-331(1980).

30 [2] Pastor N., Pinero D., Valdes A.M., Soberon X. Mol. Microbiol. 4:1957-1965(1990).

[3] Bush K. Antimicrob. Agents Chemother. 33:259-263(1989).

[4] Joris B., Ghuysen J.-M., Dive G., Renard A., Dideberg O., Charlier P., Frere J.M., Kelly J.A., Boyington J.C., Moews P.C., Knox J.R. Biochem. J. 250:313-324(1988).

920. Biotin protein ligase (BPL)

Biotin is covalently attached at the active site of certain enzymes that transfer carbon dioxide from bicarbonate to organic acids to form cellular metabolites. Biotin protein ligase (BPL) is the enzyme responsible for attaching biotin to a specific lysine at the active site of biotin enzymes. Each organism probably has only one BPL. Biotin attachment is a two step reaction that results in the formation of an amide linkage between the carboxyl group of biotin and the epsilon-amino group of the modified lysine [2].

Number of members: 26

[1] Wilson KP, Shewchuk LM, Brennan RG, Otsuka AJ, Matthews BW; Medline: 93028443 "Escherichia coli biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains." Proc Natl Acad Sci USA 1992;89:9257-9261.

[2] Chapman-Smith A, Cronan JE Jr; Medline: 10470036 "The enzymatic biotinylation of proteins: a post-translational modification of exceptional specificity." Trends Biochem Sci 1999;24:359-363.

921. (BRCA2_repeat)

The alignment covers only the most conserved region of the repeat. Respiratory-chain NADH dehydrogenase 30 Kd subunit signature

[1] Bork P, Blomberg N, Nilges M; Medline: 96241568 "Internal repeats in the BRCA2 protein sequence." Nat Genet 1996;13:22-23.

Number of members: 63

922. (C6)

This domain of unknown function is found in the C. elegans protein Swiss:Q19522. It is presumed to be an extracellular domain. The C6 domain contains six conserved cysteine

residues in most copies of the domain. However some copies of the domain are missing cysteine residues 1 and 3 suggesting that these form a disulphide bridge.

Number of members: 23

5 923. Cadherin cytoplasmic region (Cadherin_C_term)

Cadherins are vital in cell-cell adhesion during tissue differentiation. Cadherins are linked to the cytoskeleton by catenins. Catenins bind to the cytoplasmic tail of the cadherin. Cadherins cluster to form foci of homophilic binding units. A key determinant to the strength of the binding that it is mediated by cadherins is the juxtamembrane region of the cadherin. This region induces clustering and also binds to the protein p120ctn [1].

Number of members: 59

[1] Yap AS, Niessen CM, Gumbiner BM; Medline: 98234411 "The juxtamembrane region of the cadherin cytoplasmic tail supports lateral clustering, adhesive strengthening, and interaction with p120ctn." J Cell Biol 1998;141:779-789.

[2] Barth AI, Nathke IS, Nelson WJ; Medline: 97471931 "Cadherins, catenins and APC protein: interplay between cytoskeletal complexes and signaling pathways." Curr Opin Cell Biol 1997;9:683-690.

[3] Braga VM, Machesky LM, Hall A, Hotchin NA; Medline: 97327766 "The small GTPases Rho and Rac are required for the establishment of cadherin-dependent cell-cell contacts." J Cell Biol 1997;137:1421-1431.

25 924. Clathrin propeller repeat (Clathrin_propel)

Clathrin is the scaffold protein of the basket-like coat that surrounds coated vesicles. The soluble assembly unit, a triskelion, contains three heavy chains and three light chains in an extended three-legged structure. Each leg contains one heavy and one light chain. The N-terminus of the heavy chain is known as the globular domain, and is composed of seven repeats which form a beta propeller [1].

Number of members: 61

[1] ter Haar E, Musacchio A, Harrison SC, Kirchhausen T; Medline: 99043510 "Atomic structure of clathrin: a beta propeller terminal domain joins an alpha zigzag linker." Cell. 1998;95:563-573.

5 925. Respiratory-chain NADH dehydrogenase 30 Kd subunit signature (complex1_30Kd)

Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 30 Kd (in mammals) which has been found to be:

- Nuclear encoded, as a precursor form with a transit peptide in mammals, and in *Neurospora crassa*.
- Mitochondrial encoded in *Paramecium* (protein P1), and in the slime mold *Dictyostelium discoideum* (ORF 209).
- Chloroplast encoded in various higher plants (ORF 159). It is also present in bacteria:
- In the cyanobacteria *Synechocystis* strain PCC 6803 (gene *ndhJ*).
- Subunit C of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoC*).
- Subunit NQO5 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase.

This protein, in its mature form, consists of from 157 to 266 amino acid residues. The best conserved region is located in the C-terminal section and can be used as a signature pattern.

25 Consensus pattern E-R-E-x(2)-[DE]-[LIVMFY](2)-x(6)-[HK]-x(3)-[KRP]-x-[LIVM]-[LIVMYS] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT/NONE.

[1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

30 [2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

926. Respiratory-chain NADH dehydrogenase 49 Kd subunit signature (complex1_49Kd)

758

Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 49 Kd (in mammals), which is the third largest subunit of complex I and is a component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind a 4Fe-4S iron-sulfur cluster. The 49 Kd subunit has been found to be:

- Nuclear encoded, as a precursor form with a transit peptide in mammals, and in *Neurospora crassa*.

- Mitochondrial encoded in protozoan such as *Paramecium* (ORF 400), *Leishmania* and *Trypanosoma* (MURF 3).

- Chloroplast encoded in various higher plants (ORF 392).

The 49 Kd subunit is highly similar to [3,4]:

- Subunit D of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoD*).

- Subunit NQO4 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase.

- Subunit 5 of *Escherichia coli* formate hydrogenlyase (gene *hycE*).

- Subunit G of *Escherichia coli* hydrogenase-4 (gene *hyfG*).

A highly conserved region was selected as signature pattern, located in the N-terminal section of this subunit.

Consensus pattern [LIVMH]-H-[RT]-[GA]-x-E-K-[LIVMTN]-x-E-x-[KRQ] Sequences known to belong to this class detected by the patternALL.

[1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

[2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

[3] Fearnley I.M., Walker J.E. Biochim. Biophys. Acta 1140:105-134(1992).

[4] Weidner U., Geier S., Ptöck A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-122(1993).

927. (COX2)

Cytochrome c oxidase (EC 1.9.3.1) [1,2] is an oligomeric enzymatic complex which is a component of the respiratory chain and is involved in the transfer of electrons from cytochrome c to oxygen. In eukaryotes this enzyme complex is located in the mitochondrial inner membrane; in aerobic prokaryotes it is found in the plasma membrane. The enzyme complex consists of 3-4 subunits (prokaryotes) to up to 13 polypeptides (mammals).

Subunit 2 (CO II) transfers the electrons from cytochrome c to the catalytic subunit 1. It contains two adjacent transmembrane regions in its N-terminus and the major part of the protein is exposed to the periplasmic or to the mitochondrial intermembrane space, respectively. CO II provides the substrate-binding site and contains a copper center called Cu(A), probably the primary acceptor in cytochrome c oxidase. An exception is the corresponding subunit of the cbb3-type oxidase which lacks the copper A redox-center. Several bacterial CO II have a C-terminal extension that contains a covalently bound heme c.

It has been shown [3,4] that nitrous oxide reductase (EC 1.7.99.6) (gene nosZ) of *Pseudomonas* has sequence similarity in its C-terminus to CO II. This enzyme is part of the bacterial respiratory system which is activated under anaerobic conditions in the presence of nitrate or nitrous oxide. NosZ is a periplasmic homodimer that contains a dinuclear copper center, probably located in a 3-dimensional fold similar to the cupredoxin-like fold that has been suggested for the copper-binding site of CO II [3].

The dinuclear purple copper center is formed by 2 histidines and 2 cysteines [5]. This region was used as a signature pattern. The conserved valine and the conserved methionine are said to be involved in stabilizing the copper-binding fold by interacting with each other.

Consensus pattern V-x-H-x(33,40)-C-x(3)-C-x(3)-H-x(2)-M [The two C's and two H's are copper ligands] Sequences known to belong to this class detected by the patternALL, except for *Paramecium primaurelia* as well as in some plants where the pattern ends with Thr; an RNA editing event at this position could change this Thr to Met.

Note: cytochrome cbb(3) subunit 2 does not belong to this family.

[1] Capaldi R.A., Malatesta F., Darley-Usmar V.M. Biochim. Biophys. Acta 726:135-148(1983).

[2] Garcia-Horsman J.A., Barquera B., Rumbley J., Ma J., Gennis R.B. J. Bacteriol. 176:5587-5600(1994).

5 [3] van der Oost J., Lappalainen P., Musacchio A., Warne A., Lemieux L., Rumbley J., Gennis R.B., Aasa R., Pascher T., Malmstrom B.G., Saraste M. EMBO J. 11:3209-3217(1992).

[4] Zumft W.G., Dreutsch A., Loechele S., Cuypers H., Friedrich B., Schneider B. Eur. J. Biochem. 208:31-40(1992).

10

928. Cytochrome C assembly protein (CytC_asm)

This family consists of various proteins involved in cytochrome c assembly from mitochondria and bacteria; CycK from Rhizobium[3], CcmC from E. coli and Paracoccus denitrificans [2,1] and orf240 from wheat mitochondria [4]. The members of this family are probably integral membrane proteins with six predicted transmembrane helices. It has been proposed that members of this family comprise a membrane component of an ABC (ATP binding cassette) transporter complex. It is also proposed that this transporter is necessary for transport of some component needed for cytochrome c assembly. One member CycK contains a putative heme-binding motif [3], orf240 also contains a putative heme-binding motif and is a proposed ABC transporter with c-type heme as its proposed substrate [4]. However it seems unlikely that all members of this family transport heme nor c-type apocytochromes because CcmC in the putative CcmABC transporter transports neither [1].
Number of members: 67

25

[1] Page D, Pearce DA, Norris HA, Ferguson SJ; Medline: 97195802 "The Paracoccus denitrificans ccmA, B and C genes: cloning and sequencing, and analysis of the potential of their products to form a haem or apo-c-type cytochrome transporter. MICROBIOLOGY 1997;143:563-576.

30

[2] Thoeny-meyer L, Fischer F, Kunzler P, Ritz D, Hennecke H; Medline: 95362656 "Escherichia coli genes required for cytochrome c maturation." J. BACTERIOL 1995;177:4321-4326.

[3] Delgado MJ, Yeoman KH, Wu G, Vargas C, Davies A, Poole RK, Johnston AWB, Downie JA; Medline: 95394794 "Characterization of the cysHJKL genes involved in cytochrome c biogenesis and symbiotic nitrogen fixation in *Rhizobium leguminosarum*." J. BACTERIOL 1995;177:4927-4934.

5 [4] Bonnard G, Grienberger JM; Medline: 95124303 "A gene proposed to encode a transmembrane domain of an ABC transporter is expressed in wheat mitochondria." MOL. GEN. GENET 1995;246:91-99.

929. Cytochrome b559 subunits heme-binding site signature (cytochr_b559)

10

Cytochrome b559 [1] is an essential component of photosystem II complex from oxygenic photosynthetic organisms. It is an integral thylakoid membrane protein composed of two subunits, alpha (gene psbE) and beta (gene psbF), each of which contains a histidine residue located in a transmembrane region. The two histidines coordinate the heme iron of cytochrome b559.

15

The region around the heme-binding residue of both subunits is very similar and can be used as a signature pattern.

20

Consensus pattern[LIV]-x-[ST]-[LIVF]-R-[FYW]-x(2)-[IV]-H-[STGA]-[LIV]- [STGA]-[IV]-P [H is the heme iron ligand] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT NONE.

[1] Pakrasi H.B., de Ciechi P., Whitmarsh J. EMBO J. 10:1619-1627(1991).

25

930. Cytochrome b/b6 signatures (Cytochrome_b)

30

In the mitochondrion of eukaryotes and in aerobic prokaryotes, cytochrome b is a component of respiratory chain complex III (EC 1.10.2.2) - also known as the bc1 complex or ubiquinol-cytochrome c reductase. In plant chloroplasts and cyanobacteria, there is a analogous protein, cytochrome b6, a component of the plastoquinone-plastocyanin reductase (EC 1.10.99.1), also known as the b6f complex.

Cytochrome b/b6 [1,2] is an integral membrane protein of approximately 400 amino acid residues that probably has 8 transmembrane segments. In plants and cyanobacteria, cytochrome b6 consists of two subunits encoded by the *petB* and *petD* genes. The sequence of *petB* is colinear with the N-terminal part of mitochondrial cytochrome b, while *petD* corresponds to the C-terminal part. Cytochrome b/b6 non-covalently binds two heme groups, known as b562 and b566. Four conserved histidine residues are postulated to be the ligands of the iron atoms of these two heme groups.

Apart from regions around some of the histidine heme ligands, there are a few conserved regions in the sequence of b/b6. The best conserved of these regions includes an invariant P-E-W triplet which lies in the loop that separates the fifth and sixth transmembrane segments. It seems to be important for electron transfer at the ubiquinone redox site - called Qz or Qo (where o stands for outside) - located on the outer side of the membrane.

A schematic representation of the structure of cytochrome b/b6 is shown below.

```

+---Fe-b562---+ | +---Fe-b566--|+ |||
xxxxxxxxxxxxHxHxxxxxxxxxxxxHxHxxxxxxxxxxxxPEWxxxxxxxxxxxxxxxxxxxxx <-----
---Cytochrome-b-----> <---Cytochrome-b6-petB-----><---Cytochrome-
b6-petD----->

```

Two signature patterns were developed for cytochrome b/b6. The first includes the first conserved histidine of b/b6, which is a heme b562 ligand; the second includes the conserved PEW triplet.

Consensus pattern [DENQ]-x(3)-G-[FYWMQ]-x-[LIVMF]-R-x(2)-H [H is a heme b562 ligand] Sequences known to belong to this class detected by the patternALL, except for 5 sequences.

Consensus pattern P-[DE]-W-[FY]-[LFY](2) Sequences known to belong to this class detected by the patternALL, except for *Odocoileus hemionus* (mule deer) and *Paramecium tetraurelia* cytochrome b.

[1] Howell N. J. Mol. Evol. 29:157-169(1989).

[2] Esposti M.D., de Vries S., Crimi M., Ghelli A., Patarnello T., Meyer A. Biochim. Biophys. Acta 1143:243-271(1993).

5

931. Phorbol esters / diacylglycerol binding domain (DAG_PE-bind)

Diacylglycerol (DAG) is an important second messenger. Phorbol esters (PE) are analogues of DAG and potent tumor promoters that cause a variety of physiological changes when administered to both cells and tissues. DAG activates a family of serine/threonine protein kinases, collectively known as protein kinase C (PKC) [1]. Phorbol esters can directly stimulate PKC. The N- terminal region of PKC, known as C1, has been shown [2] to bind PE and DAG in a phospholipid and zinc-dependent fashion. The C1 region contains one or two copies (depending on the isozyme of PKC) of a cysteine-rich domain about 50 amino-acid residues long and essential for DAG/PE-binding. Such a domain has also been found in the following proteins:

- Diacylglycerol kinase (EC 2.7.1.107) (DGK) [3], the enzyme that converts DAG into phosphatidate. It contains two copies of the DAG/PE-binding domain in its N-terminal section. At least five different forms of DGK are known in mammals.
- N-chimaerin. A brain specific protein which shows sequence similarities with the BCR protein at its C-terminal part and contains a single copy of the DAG/PE-binding domain at its N-terminal part. It has been shown [4,5] to be able to bind phorbol esters.
- The raf/mil family of serine/threonine protein kinases. These protein kinases contain a single N-terminal copy of the DAG/PE-binding domain.
- The unc-13 protein from *Caenorhabditis elegans*. Its function is not known but it contains a copy of the DAG/PE-binding domain in its central section and has been shown to bind specifically to a phorbol ester in the presence of calcium [6].
- The vav oncogene. Vav was generated by a genetic rearrangement during gene transfer assays. Its expression seems to be restricted to cells of hematopoietic origin. Vav seems to contain a DAG/PE-binding domain in the central part of the protein.
- The *Drosophila* GTPase activating protein rotund.

10

15
20
25
30

The DAG/PE-binding domain binds two zinc ions; the ligands of these metal ions are probably the six cysteines and two histidines that are conserved in this domain. A signature pattern was developed that spans completely the DAG/PE domain.

- 5 Consensus pattern H-x-[LIVMFYW]-x(8,11)-C-x(2)-C-x(3)-[LIVMFC]-x(5,10)- C-x(2)-C-x(4)-[HD]-x(2)-C-x(5,9)-C [All the C and H are involved in binding Zinc] Sequences known to belong to this class detected by the pattern ALL, except a few DGK's.

[1] Azzi A., Boscoboinik D., Hensey C. Eur. J. Biochem. 208:547-557(1992).

- 10 [2] Ono Y., Fujii T., Igarashi K., Kuno T., Tanaka C, Kikkawa U., Nishizuka Y. Proc. Natl. Acad. Sci. U.S.A. 86:4868-4871(1989).

[3] Sakane F., Yamada K., Kanoh H., Yokoyama C., Tanabe T. Nature 344:345-348(1990).

[4] Ahmed S., Kozma R., Monfries C., Hall C., Lim H.H., Smith P., Lim L. Biochem. J. 272:767-773(1990).

[5] Ahmed S., Kozma R., Lee J., Monfries C., Harden N., Lim L. Biochem. J. 280:233-241(1991).

[6] Ahmed S., Maruyama I.N., Kozma R., Lee J., Brenner S., Lim L. Biochem. J. 287:995-999(1992).

[7] Boguski M.S., Bairoch A., Attwood T.K., Michaels G.S. Nature 358:113-113(1992).

932. 3-dehydroquinate synthase (DHQ_synthase)

[1] Barten R, Meyer TF; Medline: 98273626 "Cloning and characterisation of the Neisseria gonorrhoeae aroB gene." Mol Gen Genet 1998;258:34-44.

- 25 [2] Hawkins AR, Lamb HK; Medline: 96048023 "The molecular biology of multidomain proteins. Selected examples." Eur J Biochem 1995;232:7-18.

The 3-dehydroquinate synthase EC:4.6.1.3 domain is present in isolation in various bacterial 3-dehydroquinate synthases and also present as a domain in the pentafunctional AROM polypeptide Swiss:P07547 [2]. 3-dehydroquinate (DHQ) synthase catalyses the formation of dehydroquinate (DHQ) and orthophosphate from 3-deoxy-D-arabino heptulosonic 7 phosphate [1]. This reaction is part of the shikimate pathway which is involved in the biosynthesis of aromatic amino acids.

765

Number of members: 25

933. Dihydrofolate reductase signature (DiHfolate_red)

- 5 Dihydrofolate reductases (EC 1.5.1.3) [1] are ubiquitous enzymes which catalyze the reduction of folic acid into tetrahydrofolic acid. They can be inhibited by a number of antagonists such as trimethoprim and methotrexate which are used as antibacterial or anticancerous agents. A signature pattern was derived from a region in the N-terminal part of these enzymes, which includes a conserved Pro-Trp dipeptide; the tryptophan has been
- 10 shown [2] to be involved in the binding of substrate by the enzyme.

Consensus pattern[LVAGC]-[LIF]-G-x(4)-[LIVMF]-P-W-x(4,5)-[DE]-x(3)-[FYIV]-x(3)-[STIQ] Sequences known to belong to this class detected by the patternALL, except for type II bacterial, plasmid-encoded, dihydrofolate reductases which do not belong to the same class of enzymes.

[1] Harpers' Review of Biochemistry, Lange, Los Altos (1985).

[2] Bolin J.T., Filman D.J., Matthews D.A., Hamlin R.C., Kraut J. J. Biol. Chem. 257:13650-13662(1982).

934. (DIL)

[1] Ponting CP; Medline: 95397417 "AF-6/cno: neither a kinesin nor a myosin, but a bit of both." Trends Biochem Sci 1995;20:265-266.

Number of members: 31

935. (DNA_gyraseB_C)

DNA topoisomerase II signature (cross-reference = TOPOISOMERASE_II)

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break.

Topoisomerase II is found in phages, archaebacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaebacteria the enzyme, known as DNA gyrase, consists of two subunits (genes *gyrA* and *gyrB* [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes *parC* and *parE*). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following representation:

```

<-----About-1400-residues----->
[-----Protein 39-*-----][----Protein 52----] Phage T4
[-----gyrB-----*-----][-----gyrA-----] Prokaryote II
Archaeobacteria
[-----parE-----*-----][-----parD-----] Prokaryote IV
[-----*-----] Eukaryote and ASF

```

'*': Position of the pattern.

As a signature pattern for this family of proteins, a region was selected that contains a highly conserved pentapeptide. The pattern is located in *gyrB*, in *parE*, and in protein 39 of phage T4 topoisomerase.

Consensus pattern [LIVMA]-x-E-G-[DN]-S-A-x-[STAG] Sequences known to belong to this class detected by the pattern ALL.

[1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).

[2] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).

[3] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J. Trends Biochem. Sci. 20:156-160(1995).

936. (DNA_topoisolIV)

DNA topoisomerase II signature (cross-reference = TOPOISOMERASE_II)

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break. Topoisomerase II is found in phages, archaeobacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaeobacteria the enzyme, known as DNA gyrase, consists of two subunits (genes *gyrA* and *gyrB* [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes *parC* and *parE*). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following representation:

<-----About-1400-residues----->

[-----Protein 39-*-----][----Protein 52----] Phage T4

[-----gyrB-----*-----][-----gyrA-----] Prokaryote II Archaeobacteria

[-----parE-----*-----][-----parD-----] Prokaryote IV

[-----*-----] Eukaryote and ASF

'*': Position of the pattern.

As a signature pattern for this family of proteins, a region was selected that contains a highly conserved pentapeptide. The pattern is located in *gyrB*, in *parE*, and in protein 39 of phage T4 topoisomerase.

Consensus pattern [LIVMA]-x-E-G-[DN]-S-A-x-[STAG] Sequences known to belong to this class detected by the patternALL.

[1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).

[2] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).

[3] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J. Trends Biochem. Sci. 20:156-160(1995).

937. Prolyl oligopeptidase family serine active site (DPPIV_N_term)

5

The prolyl oligopeptidase family [1,2,3] consist of a number of evolutionary related peptidases whose catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. The known members of this family are listed below.

10 - Prolyl endopeptidase (EC 3.4.21.26) (PE) (also called post-proline cleaving enzyme). PE is an enzyme that cleaves peptide bonds on the C-terminal side of prolyl residues. The sequence of PE has been obtained from a mammalian species (pig) and from bacteria (Flavobacterium meningosepticum and Aeromonas hydrophila); there is a high degree of sequence conservation between these sequences.

15 - Escherichia coli protease II (EC 3.4.21.83) (oligopeptidase B) (gene prtB) which cleaves peptide bonds on the C-terminal side of lysyl and arginanyl residues.

- Dipeptidyl peptidase IV (EC 3.4.14.5) (DPP IV). DPP IV is an enzyme that removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline.

20 - Yeast vacuolar dipeptidyl aminopeptidase A (DPAP A) (gene: STE13) which is responsible for the proteolytic maturation of the alpha-factor precursor.

- Yeast vacuolar dipeptidyl aminopeptidase B (DPAP B) (gene: DAP2).

25 - Acylamino-acid-releasing enzyme (EC 3.4.19.1) (acyl-peptide hydrolase). This enzyme catalyzes the hydrolysis of the amino-terminal peptide bond of an N-acetylated protein to generate a N-acetylated amino acid and a protein with a free amino-terminus.

A conserved serine residue has experimentally been shown (in E.coli protease II as well as in pig and bacterial PE) to be necessary for the catalytic mechanism. This serine, which is part of the catalytic triad (Ser, His, Asp), is generally located about 150 residues away from the C-terminal extremity of these enzymes (which are all proteins that contains about 700 to 800 amino acids).

30

769

Consensus pattern D-x(3)-A-x(3)-[LIVMFYW]-x(14)-G-x-S-x-G-G-[LIVMFYW](2) [S is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for yeast DPAP A.

- 5 Note: these proteins belong to families S9A/S9B/S9C in the classification of peptidases [4,E1].

[1] Rawlings N.D., Polgar L., Barrett A.J. Biochem. J. 279:907-911(1991).

[2] Barrett A.J., Rawlings N.D. Biol. Chem. Hoppe-Seyler 373:353-360(1992).

10 [3] Polgar L., Szabo E. Biol. Chem. Hoppe-Seyler 373:361-366(1992).

[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

938. Deoxyhypusine synthase (DS)

5 Eukaryotic initiation factor 5A (eIF-5A) contains an unusual amino acid, hypusine [N epsilon-(4-aminobutyl-2-hydroxy)lysine]. The first step in the post-translational formation of hypusine is catalysed by the enzyme deoxyhypusine synthase (DS) EC:1.1.1.249. The modified version of eIF-5A, and DS, are required for eukaryotic cell proliferation [1].

10 Number of members: 9

[1] Liao DI, Wolff EC, Park MH, Davies DR; Medline: 98154315 "Crystal structure of the NAD complex of human deoxyhypusine synthase: an enzyme with a ball-and-chain mechanism for blocking the active site." Structure 1998;6:23-32.

939. (DUF21)

30 Many of the sequences in this family are annotated as hemolysins, however this is due to a similarity to Swiss:Q54318 that does not contain this domain. This domain is found in the N-terminus of the proteins adjacent to two intracellular CBS domains CBS.

Number of members: 42

940. (DUF59)

This family includes prokaryotic proteins of unknown function. The family also includes
 5 PhaH Swiss:O84984 from *Pseudomonas putida*. PhaH forms a complex with PhaF
 Swiss:O84982, PhaG Swiss:O84983 and PhaI Swiss:O84985, which hydroxylates
 phenylacetic acid to 2-hydroxyphenylacetic acid [1]. So members of this family may all be
 components of ring hydroxylating complexes.

Number of members: 15

[1] Olivera ER, Minambres B, Garcia B, Muniz C, Moreno MA, Ferrandez A, Diaz E, Garcia
 10 JL, Luengo JM; Medline: 98263372 "Molecular characterization of the phenylacetic acid
 catabolic pathway in *Pseudomonas putida* U: the phenylacetyl-CoA catabolon." Proc Natl
 Acad Sci U S A 1998;95:6419-6424.

941. (DUF82)

The protein contains four conserved cysteines that may be involved in metal binding or
 disulphide bridges.

Number of members: 4

942. Riboflavin kinase / FAD synthetase (FAD_Synth)

This family consists part of the bifunctional enzyme riboflavin kinase / FAD synthetase.

25 These enzymes have both ATP:riboflavin 5'-phospho transferase and ATP:FMN-
 adenylyltransferase activities [1]. They catalyse the 5'-phosphorylation of riboflavin to FMN
 and the adenylylation of FMN to FAD [1].

CAUTION: It is not clear if this region of the enzymes catalyses either or both of the
 enzymatic reactions.

30 Number of members: 27

[1] Manstein DJ, Pai EF; Medline: 87057286 "Purification and characterization of FAD
 synthetase from *Brevibacterium ammoniagenes*." J Biol Chem 1986;261:16169-16173.

943. [2Fe-2S] binding domain (fer2_2)

[1] Romao MJ, Archer M, Moura I, Moura JJ, LeGall J, Engh R, Schneider M, Hof P, Huber R; Medline: 96072968 "Crystal structure of the xanthine oxidase-related aldehyde oxidoreductase from *D. gigas*." Science 1995;270:1170-1176.

Number of members: 53

944. Filovirus glycoprotein (Filo_glycop)

This family includes an extracellular region from the envelope glycoprotein of Ebola and Marburg viruses. This region is also produced as a separate transcript that gives rise to a non-structural, secreted glycoprotein, which is produced in large amounts and has an unknown function [1]. Processing of this protein may be involved in viral pathogenicity [2].

Number of members: 23

[1] Volchkov VE, Feldmann H, Volchkova VA, Klenk HD; Medline: 98245155 "Processing of the Ebola virus glycoprotein by the proprotein convertase furin." Proc Natl Acad Sci U S A 1998;95:5762-5767.

[2] Sanchez A, Trappier SG, Mahy BW, Peters CJ, Nichol ST; Medline: 96195018 "The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing." Proc Natl Acad Sci U S A 1996;93:3602-3607.

945. Frataxin-like domain (Frataxin_Cyay)

This family contains proteins that have a domain related to the globular C-terminus of Frataxin the protein that is mutated in Friedreich's ataxia. This domain is found in a family of bacterial proteins. The function of this domain is currently unknown.

Number of members: 12

[1] Gibson TJ, Koonin EV, Musco G, Pastore A, Bork P; Medline: 97084946 "Friedreich's ataxia protein: phylogenetic evidence for mitochondrial dysfunction." Trends Neurosci 1996;19:465-468.

946. (GAF)

Domain present in phytochromes and cGMP-specific phosphodiesterases.

5 Number of members: 296

[1] Aravind L, Ponting CP; Medline: 98094688 "The GAF domain: an evolutionary link between diverse phototransducing proteins." Trends Biochem Sci 1997;22:458-459.

10 947. Galaptin signature (Gal-bind_lectin)

All vertebrates synthesize soluble galactoside-binding lectins [1,2,3] (also known as galectins, galaptins or S-lectin). These carbohydrate-binding proteins are developmentally regulated. Although their exact physiological role is not yet clear they seem to be involved in differentiation, cellular regulation and tissue construction. The sequence of galactoside-binding lectins from electric eel (electrolectin), conger eel (congerin), chicken and a number of mammalian species is known. These lectins are proteins of about 130 to 140 amino acid residues (14 Kd to 16 Kd).

20 A number of other proteins are known to belong to this family:

- Galectin-3 (also known as MAC-2 antigen; CBP-35 or IgE-binding protein), a 35 Kd lectin which binds immunoglobulin E and which is composed of two domains: a N-terminal domain that consist of tandem repeats of a glycine/ proline-rich sequence and a C-terminal galaptin domain.

25 - Galectin-4 [4], which is composed of two galaptin domains.

- Galectin-5.

- Galectin-7 [5], a keratinocyte protein which could be involved in cell-cell and/or cell-matrix interactions necessary for normal growth control.

- Galectin-8 [6], which is composed of two galaptin domains.

30 - Galectin-9 [7], which is composed of two galaptin domains.

- Human eosinophil lysophospholipase (EC 3.1.1.5) [8] (Charcot-Leyden crystal protein), a protein that may have both an enzymatic and a lectin activities. It forms hexagonal

bipyramidal crystals in tissues and secretions from sites of eosinophil-associated inflammation.

- *Caenorhabditis elegans* 32 Kd lactose-binding lectin [9]. This lectin is composed of two galactin domains.

5 - *Caenorhabditis elegans* lec-7 and lec-8.

One of the conserved regions of these lectins contains a tryptophan that has been shown [10] to be essential to the binding of galactosides. This region was used as a signature pattern for these proteins.

10 Consensus pattern W-[GEK]-x-[EQ]-x-[KRE]-x(3,6)-[PCTF]-[LIVMF]-[NQE GSKV]-x-[GH]-x(3)-[DENKHS]-[LIVMFC] [W binds carbohydrate] Sequences known to belong to this class detected by the pattern ALL, except for pig galectin 4.

[1] Barondes S.H., Gitt M.A., Leffler H., Cooper D.N.W. *Biochimie* 70:1627-1632(1988).

15 [2] Hirabayashi J., Kasai K.-I. *J. Biochem.* 104:1-4(1988).

[3] Barondes S.H., Castronovo V., Cooper D.N.W., Cummings R.D., Drickamer K., Feizi T., Gitt M.A., Hirabayashi J., Hughes C., Kasai K.-I., Leffler H., Liu F.-T., Lotan R., Mercurio A.M., Monsigny M., Pillair S., Poirer F., Raz A., Rigby P.W.J., Rini J.M., Wang J.L. *Cell* 76:597-598(1994).

20 [4] Oda Y., Herrmann J., Gitt M., Turck C.W., Burlingame A.L., Barondes S.H., Leffler H. *J. Biol. Chem.* 268:5929-5939(1993).

[5] Madsen P., Rasmussen H.H., Flint T., Gromov P., Kruse T.A., Honore B., Vorum H., Celis J.E. *J. Biol. Chem.* 270:5823-5829(1995).

[6] Hadari Y.R., Paz K., Dekel R., Mestrovic T., Accili D., Zick Y. *J. Biol. Chem.* 270:3447-25 3453(1995).

[7] Wada J., Kanwar Y.S. *J. Biol. Chem.* 272:6078-6086(1997).

[8] Ackerman S.J., Corrette S.E., Rosenberg H.F., Bennett J.C., Mastrianni D.M., Nicholson-Weller A., Weller P.F., Chin D.T., Tenen D.G. *J. Immunol.* 150:456-468(1993).

[9] Hirabayashi J., Satoh M., Kasai K.-I. *J. Biol. Chem.* 267:15485-15490(1992).

30 [10] Abbott W.M., Feizi T. *J. Biol. Chem.* 266:5552-5557(1991).

948. (GARS) Phosphoribosylglycinamide synthetase signature (phosphoribosylamine glycine ligase)

PROSITE: PDOC00164; cross-reference(s): PS00184

[1] catalyzes the second step in the de novo biosynthesis of purine, the ATP-dependent addition of 5-phosphoribosylamine to glycine to form 5'phosphoribosylglycinamide.

5 In bacteria GARS is a monofunctional enzyme (encoded by the purD gene), in of a bifunctional enzyme (encoded by the ADE5,7 gene), in higher eukaryotes it is part, with AIRS and with phosphoribosylglycinamide formyltransferase (GART) of a trifunctional enzyme (GARS-AIRS-GART).

10 The sequence of GARS is well conserved. A highly conserved octapeptide was selected as a signature pattern.

Consensus pattern R-F-G-D-P-E-x-[QM]

Sequences known to belong to this class detected by the pattern ALL.

5 [1] Aiba A., Mizobuchi K. J. Biol. Chem. 264:21239-21246(1989).

949. GLTT - GLTT repeat (12 copies)

This short repeat of unknown function is found in multiple copies in several C. elegans proteins. The repeat is five residues long and consists of XGLTT where X can be any amino acid. Number of members: 34.

950. Glu_synthase - Conserved region in glutamate synthase

25 This family represents a region of the glutamate synthase protein. This region is expressed as a separate subunit in the glutamate synthase alpha subunit from archaeobacteria, or part of a large multidomain enzyme in other organisms. The aligned region of these proteins contains a putative FMN binding site and Fe-S cluster. Number of members: 44.

30 [1] Medline: 97082505. Sequence of the GLT1 gene from Saccharomyces cerevisiae reveals the domain structure of yeast glutamate synthase. Filetici P, Martegani MP, Valenzuela L, Gonzalez A, Ballario P; Yeast 1996;12:1359-1366.

951. (Glyco_hydro_2) Glycosyl hydrolases family 2 signatures

GLYCOSYL_HYDROL_F2_1; PS00608; GLYCOSYL_HYDROL_F2_2

775

It has been shown [1,2,E1] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

-Beta-galactosidases (EC 3.2.1.23) from bacteria such as *Escherichia coli* (genes *lacZ* and *ebgA*), *Clostridium acetobutylicum*, *Clostridium thermosulfurogenes*, *Klebsiella pneumoniae*, *Lactobacillus delbrueckii*, or *Streptococcus thermophilus* and from the fungi *Kluyveromyces lactis*.

-Beta-glucuronidase (EC 3.2.1.31) from *Escherichia coli* (gene *uidA*) and from mammals. One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [3], in *Escherichia coli lacZ*, to be the general acid/base catalyst in the active site of the enzyme. This region has been used as a signature pattern. A highly conserved region located some sixty residues upstream from the active site glutamate has been selected as a second signature pattern.

Consensus pattern N-x-[LIVMFYWD]-R-[STACN](2)-H-Y-P-x(4)-[LIVMFYWS](2)-x(3)-[DN]-x(2)-G-[LIVMFYW](4) Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [DENQLF]-[KRVW]-N-[HRY]-[STAPPV]-[SAC]-[LIVMFS](3)-W-[GS]-x(2,3)-N-E [E is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for *Rhizobium meliloti lacZ*.

[1]Henrissat B. *Biochem. J.* 280:309-316(1991).

[2]Schroeder C.J., Robert C., Lenzen G., McKay L.L., Mercenier A. J. *Gen. Microbiol.* 137:369-380(1991).

[3]Gebler J.C., Aebersold R., Withers S.G. *J. Biol. Chem.* 267:11126-11130(1992).

952. (Glyco_hydro_3) Glycosyl hydrolases family 3 active site

PROSITE: PDOC00621. PROSITE cross-reference(s)PS00775; GLYCOSYL_HYDROL_F3

It has been shown [1,2] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

-Beta glucosidases (EC 3.2.1.21) from the fungi *Aspergillus wentii* (A-3), *Hansenula anomala*, *Kluyveromyces fragilis*, *Saccharomycopsis fibuligera*, (BGL1 and BGL2), *Schizophyllum commune* and *Trichoderma reesei* (BGL1).

-Beta glucosidases from the bacteria *Agrobacterium tumefaciens* (Cbg1), *Butyrivibrio fibrisolvens* (bglA), *Clostridium thermocellum* (bglB), *Escherichia coli* (bglX), *Erwinia chrysanthemi* (bgxA) and *Ruminococcus albus*.

-*Alteromonas* strain O-7 beta-hexosaminidase A (EC 3.2.1.52).

5 -*Bacillus subtilis* hypothetical protein yzbA.

-*Escherichia coli* hypothetical protein ycfO and HI0959, the corresponding *Haemophilus influenzae* protein.

One of the conserved regions in these enzymes is centered on a conserved aspartic acid residue which has been shown [3], in *Aspergillus wentii* beta-glucosidase A3, to be
10 implicated in the catalytic mechanism. This region was used as a signature pattern.

Consensus pattern[LIVM](2)-[KR]-x-[EQK]-x(4)-G-[LIVMFT]-[LIVT]-[LIVMF]-[ST]-D-x(2)-[SGADNI] [D is the active site residue]

Sequences known to belong to this class detected by the patternALL.

15 [1]Henrissat B. Biochem. J. 280:309-316(1991).

[2]Castle L.A., Smith K.D., Morris R.O. J. Bacteriol. 174:1478-1486(1992).

[3]Bause E., Legler G. Biochim. Biophys. Acta 626:459-465(1980).

20 953. GP120 - Envelope glycoprotein GP120

The entry of HIV requires interaction of viral GP120 with Swiss:P01730 and a chemokine receptor on the cell surface. Number of members: 17891

[1]Medline: 98303379. Structure of an HIV gp120 envelope glycoprotein in complex with
25 the CD4 receptor and a neutralizing human antibody. Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA; Nature 1998;393:648-659.

954. (GSPII_E) Bacterial type II secretion system protein E signature

PROSITE: PDOC00567. PROSITE cross-reference(s) PS00662; T2SP_E

30 A number of bacterial proteins, some of which are involved in a general secretion pathway (GSP) for the export of proteins (also called the type II pathway) [1,2], have been found to be evolutionary related. These proteins are listed below:

-The 'E' protein from the GSP operon of: *Aeromonas* (gene *exeE*); *Erwinia* (gene *outE*); *Escherichia coli* (gene *yheG*); *Klebsiella pneumoniae* (gene *pulE*); *Pseudomonas aeruginosa* (gene *xcpR*); *Vibrio cholerae* (gene *epsE*) and *Xanthomonas campestris* (gene *xpsE*).

-*Agrobacterium tumefaciens* Ti plasmid *virB* operon protein 11. This protein is required for the transfer of T-DNA to plants.

-*Bacillus subtilis* *comG* operon protein 1 which is required for the uptake of DNA by competent *Bacillus subtilis* cells.

-*Aeromonas hydrophila* *tapB*, involved in type IV pilus assembly.

-*Pseudomonas* protein *pilB*, which is essential for the formation of the pili.

-*Pseudomonas aeruginosa* protein twitching mobility protein *pilT*.

-*Neisseria gonorrhoeae* type IV pilus assembly protein *pilF*.

-*Vibrio cholerae* protein *tcpT*, which is involved in the biosynthesis of the *tcp* pilus.

-*Escherichia coli* protein *hofB* (*hopB*).

-*Escherichia coli* hypothetical protein *ygcB*.

-*Escherichia coli* hypothetical protein *yggR*.

These proteins have from 344 (*pilT* and *virB11*) to 568 (*tapB*) amino acids, they are probably cytoplasmically located and, on the basis of the presence of a conserved P-loop region (see <PDOC00017>), probably bind ATP. A region that overlaps the 'B' motif of ATP-binding proteins was selected as a signature pattern.

Consensus pattern[LIVM]-R-x(2)-P-D-x-[LIVM](3)-G-E-[LIVM]-R-D

Sequences known to belong to this class detected by the patternALL, except for *ygcB*.

[1]Salmond G.P.C., Reeves P.J. Trends Biochem. Sci. 18:7-12(1993).

[2]Hobbs M., Mattick J.S. Mol. Microbiol. 10:233-243(1993).

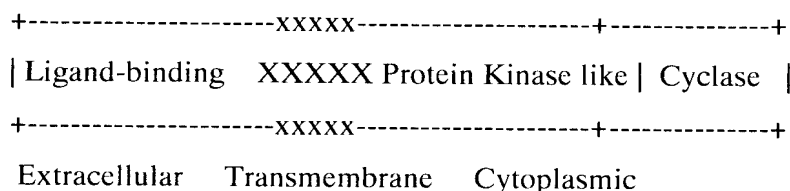
955. (guanylate_cyc) Guanylate cyclases signature

PROSITE: PDOC00425. PROSITE cross-reference(s) PS00452;

GUANYLATE_CYCLASES Guanylate cyclases (EC 4.6.1.2) [1 to 4] catalyze the formation of cyclic GMP (cGMP) from GTP. cGMP acts as an intracellular messenger, activating cGMP dependent kinases and regulating CGMP-sensitive ion channels. The role of cGMP as a second messenger in vascular smooth muscle relaxation and retinal photo-

transduction is well established. Guanylate cyclase is found both in the soluble and particular fraction of eukaryotic cells. The soluble and plasma membrane-bound forms differ in structure, regulation and other properties.

Most currently known plasma membrane-bound forms are receptors for small polypeptides. The topology of such proteins is the following: they have a N-terminal extracellular domain which acts as the ligand binding region, then a transmembrane domain, followed by a large cytoplasmic C-terminal region that can be subdivided into two domains: a protein kinase-like domain that appears important for proper signalling and a cyclase catalytic domain. This topology is schematically represented below.



The known guanylate cyclase receptors are:

- The sea-urchins receptors for speract and resact, which are small peptides that stimulate sperm motility and metabolism.
- The receptors for natriuretic peptides (ANF). Two forms of ANF receptors with guanylate cyclase activity are currently known: GC-A (or ANP-A) which seems specific to atrial natriuretic peptide (ANP), and GC-B (or ANP-B) which seems to be stimulated more effectively by brain natriuretic peptide (BNP) than by ANP.
- The receptor for Escherichia coli heat-stable enterotoxin (GC-C). The endogenous ligand for this intestinal receptor seems to be a small peptide called guanylin.
- Retinal guanylate cyclase (retGC) which probably plays a specific functional role in the rods and/or cones of photoreceptors. It is not known if this protein acts as receptor, but its structure is similar to that of the other plasma membrane-bound GCs.

The soluble forms of guanylate cyclase are cytoplasmic heterodimers. The two subunits, alpha and beta are proteins of from 70 to 82 Kd which are highly related. Two forms of beta subunits are currently known: beta-1 which seems to be expressed in lung and brain, and beta-2 which is more abundant in kidney and liver.

The membrane and cytoplasmic forms of guanylate cyclase share a conserved domain which is probably important for the catalytic activity of the enzyme. Such a domain is also

found twice in the different forms of membrane-bound adenylate cyclases (also known as class-III) [5,6] from mammals, slime mold or *Drosophila*. A consensus pattern was derived from the most conserved region in that domain.

5 Consensus pattern G-V-[LIVM]-x(0,1)-G-x(5)-[FY]-x-[LIVM]-[FYW]-[GS]-[DNTHKW]-[DNT]-[IV]-[DNTA]-x(5)-[DE]

Sequences known to belong to this class detected by the pattern ALL, except for the sea urchin *Arbacia punctulata* resact receptor which lack this domain.

Note this pattern will detect both domains of adenylate cyclases class-III.

10

[1] Koesling D., Boehme E., Schultz G. FASEB J. 5:2785-2791(1991).

[2] Garbers D.L. New Biol. 2:499-504(1990).

[3] Garbers D.L. Cell 71:1-4(1992).

[4] Yuen P.S.T., Garbers D.L. Annu. Rev. Neurosci. 15:193-225(1992).

5 [5] Iyengar R. FASEB J. 7:768-775(1993).

[6] Barzu O., Danchin A. Prog. Nucleic Acid Res. Mol. Biol. 49:241-283(1994).

956. Hemolysin-type calcium-binding region signature (HemolysinCabinD)

20

Gram-negative bacteria produce a number of proteins which are secreted into the growth medium by a mechanism that does not require a cleaved N-terminal signal sequence. These proteins, while having different functions, seem [1] to share two properties: they bind calcium and they contain a variable number of tandem repeats consisting of a nine amino acid motif rich in glycine, aspartic acid and asparagine. It has been shown [2] that such a domain is involved in the binding of calcium ions in a parallel beta roll structure. The proteins which are currently known to belong to this category are:

25

- Hemolysins from various species of bacteria. Bacterial hemolysins are exotoxins that attack blood cell membranes and cause cell rupture. The hemolysins which are known to contain such a domain are those from: *E. coli* (gene hlyA), *A. pleuropneumoniae* (gene appA), *A.*

30

- actinomycetemcomitans* and *P. haemolytica* (leukotoxin) (gene lktA).

- Cyclolysin from *Bordetella pertussis* (gene cyaA). A multifunctional protein which is both an adenylate cyclase and a hemolysin.

780

- Extracellular zinc proteases: serralyisin (EC 3.4.24.40) from *Serratia*, prtB and prtC from *Erwinia chrysanthemi* and aprA from *Pseudomonas aeruginosa*.

- Nodulation protein nodO from *Rhizobium leguminosarum*.

A signature pattern was derived from conserved positions in the sequence of the calcium-binding domain.

Consensus pattern D-x-[LI]-x(4)-G-x-D-x-[LI]-x-G-G-x(3)-D Sequences known to belong to this class detected by the pattern ALL.

Note: This pattern is found once in nodO and the extracellular proteases but up to 5 times in some hemolysin/cyclolysins.

[1] Economou A., Hamilton W.D.O., Johnston A.W.B., Downie J.A. EMBO J. 9:349-354(1990).

[2] Baumann U., Wu S., Flaherty K.M., McKay D.B. EMBO J. 12:3357-3364(1993).

957. Hint module (Hint)

This is an alignment of the Hint module in the Hedgehog proteins. It does not include any Inteins which also possess the Hint module.

Number of members: 36

[1] Hall TM, Porter JA, Young KE, Koonin EV, Beachy PA, Leahy DJ; Medline: 97474313 "Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins." Cell 1997;91:85-97.

958. Hydantoinase/oxoprolinase (Hydantoinase)

This family includes the enzymes hydantoinase and oxoprolinase EC:3.5.2.9. Both reactions involve the hydrolysis of 5-membered rings via hydrolysis of their internal imide bonds [1].

Number of members: 14

[1] Ye GJ, Breslow EB, Meister A, Guo-jie GE\$[corrected to Ye GJ]; Medline: 97113037
“The amino acid sequence of rat kidney 5-oxo-L-prolinase determined by cDNA cloning”
[published erratum appears in J Biol Chem 1997 Feb 14;272(7):4646] J Biol Chem
1996;271:32293-32300.

5

959. IMP dehydrogenase / GMP reductase signature (IMPDH_N)

IMP dehydrogenase (EC 1.1.1.205) (IMPDH) catalyzes the rate-limiting reaction of de novo
GTP biosynthesis, the NAD-dependent reduction of IMP into XMP [1]. Inhibition of IMP
dehydrogenase activity results in the cessation of DNA synthesis. As IMP dehydrogenase is
associated with cell proliferation, it is a possible target for cancer chemotherapy. Mammalian
and bacterial IMPDHs are tetramers of identical chains. There are two IMP dehydrogenase
isozymes in humans [2].

10

GMP reductase (EC 1.6.6.8) catalyzes the irreversible and NADPH-dependent reductive
deamination of GMP into IMP [3]. It converts nucleobase, nucleoside and nucleotide
derivatives of G to A nucleotides, and maintains intracellular balance of A and G nucleotides.

5

IMP dehydrogenase and GMP reductase share many regions of sequence similarity. One of
these regions is centered on a cysteine residue thought [3] to be involved in binding IMP.
This region was used as a signature pattern.

20

Consensus pattern[LIVM]-[RK]-[LIVM]-G-[LIVM]-G-x-G-S-[LIVM]-C-x-T [C is the
putative IMP-binding residue] Sequences known to belong to this class detected by the
pattern ALL.

25

[1] Collart F.R., Huberman E. J. Biol. Chem. 263:15769-15772(1988).

[2] Natsumeda Y., Ohno S., Kawasaki H., Konno Y., Weber G., Suzuki K. J. Biol. Chem.
265:5292-5295(1990).

30

[3] Andrews S.C., Guest J.R. Biochem. J. 255:35-43(1988).

960. impB/mucB/samB family (IMS)

782

These proteins are involved in UV protection (Swiss).

Number of members: 38

961. Type II intron maturase (Intron_maturas2)

5

Group II introns use intron-encoded reverse transcriptase, maturase and DNA endonuclease activities for site-specific insertion into DNA [2]. Although this type of intron is self splicing in vitro they require a maturase protein for

10

splicing in vivo. It has been shown that a specific region of the aI2 intron is needed for the maturase function [1]. This region was found to be conserved in group II introns and called domain X [3].

Number of members: 335

[1] Moran JV, Mecklenburg KL, Sass P, Belcher SM, Mahnke D, Lewin A, Perlman P;

15

Medline: 94301788 "Splicing defective mutants of the COXI gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron aI2. Nucleic Acids Res 1994;22:2057-2064.

[2] Guo H, Zimmerly S, Perlman PS, Lambowitz AM; Medline: 98031910 "Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA." EMBO J 1997;16:6835-6848.

20

[3] Mohr G, Perlman PS, Lambowitz AM; Medline: 94077696 "Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function." Nucleic Acids Res 1993;21:4991-4997.

25

962. LAGLIDADG endonuclease (Intron_maturase)

[1] Heath PJ, Stephens KM, Monnat RJ Jr, Stoddard BL; Medline: 97331323 "The structure of I-Crel, a group I intron-encoded homing endonuclease." Nat Struct Biol 1997;4:468-476.

30

[2] Belfort M, Roberts RJ; Medline: 97402526 "Homing endonucleases: keeping the house in order." Nucleic Acids Res 1997;25:3379-3388.

[3] Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS; Medline: 98026854 "Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases

and identification of an intein that encodes a site-specific endonuclease of the HNH family.”
Nucleic Acids Res 1997;25:4626-4638.

Number of members: 220

5

963. Isopentenyl transferase (IPT)

Isopentenyl transferase / dimethylallyl transferase synthesizes isopentenyladenosine 5'-
monophosphate, a cytokinin that induces shoot formation on host plants infected with the Ti
plasmid [1].

10

Number of members: 16

[1] Canaday J, Gerad JC, Crouzet P, Otten L; Medline: 93101133 "Organization and
functional analysis of three T-DNAs from the vitopine Ti plasmid pTiS4." Mol Gen Genet
1992;235:292-303.

15

964. Laminin EGF-like (Domains III and V) (laminin_EGF)

This family is like EGF but has 8 conserved cysteines instead of 6.

Number of members: 501

20

[1] Engel J; Medline: 93041759 "Laminins and other strange proteins." Biochemistry
1992;31:10643-10651.

25

965. Legume lectins signatures (lectin_legA)

Leguminous plants synthesize sugar-binding proteins which are called legume lectins [1,2].
These lectins are generally found in the seeds. The exact function of legume lectins is not
known but they may be involved in the attachment of nitrogen-fixing bacteria to legumes and
in the protection against pathogens. Legume lectins bind calcium and manganese (or other
transition metals).

30

784

Legume lectins are synthesized as precursor proteins of about 230 to 260 amino acid residues. Some legume lectins are proteolytically processed to produce two chains: beta (which corresponds to the N-terminal) and alpha (C-terminal). The lectin concanavalin A (conA) from jack bean is exceptional in that the two chains are transposed and ligated (by formation of a new peptide bond). The N-terminus of mature conA thus corresponds to that of the alpha chain and the C-terminus to the beta chain.

Two signature patterns were developed specific to legume lectins: the first is located in the C-terminal section of the beta chain and contains a conserved aspartic acid residue important for the binding of calcium and manganese; the second one is located in the N-terminal of the alpha chain.

Consensus pattern [LIV]-[STAG]-V-[DEQV]-[FLI]-D-[ST] [D binds manganese and calcium] Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [LIV]-x-[EDQ]-[FYWKR]-V-x-[LIVF]-G-[LF]-[ST] Sequences known to belong to this class detected by the pattern ALL.

[1] Sharon N., Lis H. FASEB J. 4:3198-320(1990).

[2] Lis H., Sharon N. Annu. Rev. Biochem. 55:33-37(1986).

966. Malate synthase signature (malate_synthase)

Malate synthase (EC 4.1.3.2) catalyzes the aldol condensation of glyoxylate with acetyl-CoA to form malate - the second step of the glyoxylate bypass, an alternative to the tricarboxylic acid cycle in bacteria, fungi and plants. Malate synthase is a protein of 530 to 570 amino acids whose sequence is highly conserved across species [1]. As a signature pattern, a very conserved region was selected in the central section of the enzyme.

Consensus pattern[KR]-[DENQ]-H-x(2)-G-L-N-x-G-x-W-D-Y-[LIVM]-F Sequences known to belong to this class detected by the pattern ALL.

[1] Bruinenberg P.G., Blaauw M., Kazemier B., Ab G. Yeast 6:245-254(1990).

967. MatK/TrnK amino terminal region (MatK_N)

[1] Mohr G, Perlman PS, Lambowitz AM; Medline: 94077696 "Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function." Nucleic Acids Res 1993;21:4991-4997.

Number of members: 495

968. MOZ/SAS family (MOZ_SAS)

This region of these proteins has been suggested to be homologous to acetyltransferases [1]. However the similarity is not supported by standard sequence analysis.

Number of members: 15

[1] Kamine J, Elangovan B, Subramanian T, Coleman D, Chinnadurai G; Medline: 96182937 "Identification of a cellular protein that specifically interacts with the essential cysteine region of the HIV-1 Tat transactivator." Virology 1996;216:357-366.

[2] Reifsnyder C, Lowell J, Clarke A, Pillus L; Medline: 96376969 "Yeast SAS silencing genes and human genes associated with AML and HIV-1 Tat interactions are homologous with acetyltransferases" [see comments] [published erratum appears in Nat Genet 1997 May;16(1):109] Nat Genet 1996;14:42-49.

969. mRNA capping enzyme (mRNA_cap_enzyme)

[1] Hakansson K, Doherty AJ, Shuman S, Wigley DB; Medline: 97304383 "X-ray crystallography reveals a large conformational change during guanylyl transfer by mRNA capping enzymes." Cell 1997;89:545-553.

Number of members: 7

970. DNA mismatch repair proteins mutS family signature (MutS_C)

Mismatch repair contributes to the overall fidelity of DNA replication [1]. It involves the correction of mismatched base pairs that have been missed by the proofreading element of the DNA polymerase complex. The sequence of some proteins involved in mismatch repair in different organisms have been found to be evolutionary related [2,3]. One of these families is called mutS [4,E1], it consists of:

- Prokaryotic protein mutS protein (also called hexA in *Streptococcus pneumoniae*). Muts is thought to carry out the mismatch recognition step of DNA repair.

- Eukaryotic MSH1, which is involved in mitochondrial DNA repair.

- Eukaryotic MSH2, which is involved in nuclear postreplication mismatch repair. MSH2 heterodimerizes with MSH6. In man, MSH2 is involved in a form of familial hereditary nonpolyposis colon cancer (HNPCC).

- Eukaryotic MSH3, which is probably involved in the repair of large loops.

- Eukaryotic MSH4, which is involved in meiotic recombination.

- Eukaryotic MSH5, which is involved in meiotic recombination.

- Eukaryotic MSH6 (also known as G/T mismatch binding protein), a DNA-repair protein that binds to G/T mismatches through heterodimerization with MSH2.

- Prokaryotic protein mutS2 whose function is not yet known.

- A coral (*Sarcophyton glaucum*) mitochondrial encoded mutS-like protein.

As a signature pattern for this class of mismatch repair proteins a region rich in glycine and negatively charged residues was selected. This region is found in the C-terminal section of these proteins; about 80 residues to the C-terminal of an ATP-binding site motif 'A' (P-loop) (see <PDOC00017>).

Consensus pattern[ST]-[LIVMF]-x-[LIVM]-x-D-E-[LIVMFY]-[GC]-[RKH]-G-[GST]- x(4)-

G Sequences known to belong to this class detected by the pattern ALL, except for mutS2.

[1] Modrich P. Annu. Rev. Biochem. 56:435-466(1987).

[2] Haber L.T., Walker G.C. EMBO J. 10:2707-2715(1991).

[3] New L., Liu K., Crouse G.F. Mol. Gen. Genet. 239:97-108(1993).

[4] Eisen J.A. Nucleic Acids Res. 26:4291-4300(1998).

971. MutS family, N-terminal putative DNA binding domain (MutS_N)

This family consists of the N-terminal region of proteins in the mutS family of DNA mismatch repair proteins and is found associated with MutS_C located in the C-terminal region. The mutS family of proteins is named after the salmonella typhimurium MutS protein involved in mismatch repair; other members of the family included the eukaryotic MSH 1,2,3,4,5 and 6 proteins. These have various roles in DNA repair and recombination. Human MSH has been implicated in non-polyposis colorectal carcinoma (HNPCC) and is a mismatch binding protein [2]. The aligned region corresponds in part with domains A1, A2 (which may bind DNA) and B (which binds dsDNA in vitro) from T. thermophilus MutS as characterised in [1].

Number of members: 43

972. Domain in Myosin and Kinesin Tails (MyTH4)

Domain present twice in myosin-VIIa, and also present in 3 other myosins.

[1] Chen ZY, Hasson T, Kelley PM, Schwender BJ, Schwartz MF, Ramakrishnan M, Kimberling WJ, Mooseker MS, Corey DP; Medline: 97038686 "Molecular cloning and domain structure of human myosin-VIIa, the gene product defective in Usher syndrome 1B." Genomics 1996;36:440-448.

Number of members: 21

973. Sodium and potassium ATPases beta subunits signatures (Na_K-ATPase)

The sodium pump (Na⁺,K⁺ ATPase), located in the plasma membrane of all animal cells [1], is an heterotrimer of a catalytic subunit (alpha chain), a glycoprotein subunit of about 34 Kd (beta chain) and a small hydrophobic protein of about 6 Kd. The beta subunit seems [2] to regulate, through the assembly of alpha/beta heterodimers, the number of sodium pumps transported to the plasma membrane.

Structurally the beta subunit is composed of a charged cytoplasmic domain of about 35 residues, followed by a transmembrane region, and a large extracellular domain that contains

three disulfide bonds and glycosylation sites. This structure is schematically represented in the figure below.

+----+ +---+ +-----+ |||||

xxxxxxxxxxxxxxxxxxxxxxxxCxxxxCxCxxCxxxxxxxxCxxxxxxxxxxxxCxxxx

5 ***** <-Cyt-><TM><-----Extracellular----->

'C': conserved cysteine involved in a disulfide bond.

'*': position of the patterns.

10 Two isoforms of the beta subunit (beta-1 and beta-2) are currently known; they share about 50% sequence identity. Gastric (K⁺, H⁺) ATPase (proton pump) responsible for acid production in the stomach consist of two subunits [3]; the beta chain is highly similar to the sodium pump beta subunits. Two signature patterns were developed for beta subunits. The first is located in the cytoplasmic domain, while the second is found in the extracellular domain and contains two of the cysteines involved in disulfide bonds.

Consensus pattern [FYW]-x(2)-[FYW]-x-[FYW]-[DN]-x(6)-[LIVM]-G-R-T-x(3)-W

Sequences known to belong to this class detected by the pattern ALL.

20 Consensus pattern [RK]-x(2)-C-[RKQWI]-x(5)-L-x(2)-C-[SA]-G [The two C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern ALL, except for the beta subunit of the sodium pump of brine shrimp whose sequence is highly divergent in that region.

- 25 [1] Horisberger J.D., Lemas V., Krahenbul J.P., Rossier B.C. Annu. Rev. Physiol. 53:565-584(1991).
- [2] McDonough A.A., Gerring K., Farley R.A. FASEB J. 4:1598-1605(1990).
- [3] Toh B.-H., Gleeson P.A., Simpson R.J., Moritz R.L., Callaghan J.M., Goldkorn I., Jones C.M., Martinelli T.M., Mu F.-T., Humphris D.C., Pettitt J.M., Mori Y., Masuda T.,
- 30 Sobieszczuk P., Weinstock J., Mantamadiotis T., Baldwin G.S. Proc. Natl. Acad. Sci. U.S.A. 87:6418-6422(1990).

Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there are fifteen which are located in the membrane part, seven of which are encoded by the mitochondrial and chloroplast genomes of most species. The most conserved of these organelle-encoded subunits is known as subunit 1 (gene ND1 in mitochondrion, and NDH1 in chloroplast) and seems to contain the ubiquinone binding site.

The ND1 subunit is highly similar to subunit 4 of *Escherichia coli* formate hydrogenlyase (gene hycD), subunit C of hydrogenase-4 (gene hycC). *Paracoccus denitrificans* NQO8 and *Escherichia coli* nuoH NADH-ubiquinone oxidoreductase subunits also belong to this family [3]. Two signature patterns were developed based on conserved regions of this subunit.

Consensus pattern G-[LIVMFYKRS]-[LIVMAGP]-Q-x-[LIVMFY]-x-D-[AGIM]-[LIVMFTA]-K-[LVMYST]-[LIVMFYG]-x-[KR]-[EQG] Sequences known to belong to this class detected by the pattern ALL, except for watermelon and *Leishmania* ND1.

Consensus pattern P-F-D-[LIVMFYQ]-[STAGPVM]-E-[GAC]-E-x-[EQ]-[LIVMS]-x(2)-G Sequences known to belong to this class detected by the pattern ALL, except for *Chlamydomonas reinhardtii* and *Pisaster ochraceus* ND1, and tobacco NDH1.

[1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

[2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

[3] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-122(1993).

975. Nickel-dependent hydrogenases large subunit signatures (NiFeSe_Hases)

Hydrogenases are enzymes that catalyze the reversible activation of hydrogen and which occur widely in prokaryotes as well as in some eukaryotes. There are various types of hydrogenases, but all of them seem to contain at least one iron-sulfur cluster. They can be

broadly divided into two groups: hydrogenases containing nickel and, in some cases, also selenium (the [NiFe] and [NiFeSe] hydrogenases) and those lacking nickel (the [Fe] hydrogenases).

5 The [NiFe] and [NiFeSe] hydrogenases are heterodimer that consist of a small subunit that contains a signal peptide and a large subunit. All the known large subunits seem to be evolutionary related [1]; they contain two Cys-x-x- Cys motifs; one at their N-terminal end; the other at their C-terminal end. These four cysteines are involved in the binding of nickel [2]. In the [NiFeSe] hydrogenases the first cysteine of the C-terminal motif is a
10 selenocysteine which has experimentally been shown to be a nickel ligand [3]. Two patterns were developed which are centered on the Cys-x-x-Cys motifs.

Alcaligenes eutrophus possess a NAD-reducing cytoplasmic hydrogenase (hoxS) [4]; this enzyme is composed of four subunits. Two of these subunits (beta and delta) are responsible
15 for the hydrogenase reaction and are evolutionary related to the large and small subunits of membrane-bound hydrogenases. The alpha subunit of coenzyme F420 hydrogenase (EC 1.12.99.1) (FRH) from archaebacterial methanogens also belongs to this family.

Consensus pattern R-G-[LIVMF]-E-x(15)-[QESM]-R-x-C-G-[LIVM]-C [The two C's are
20 nickel ligands] Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [FY]-D-P-C-[LIM]-[ASG]-C-x(2,3)-H [The two C's are nickel ligands]
Sequences known to belong to this class detected by the pattern ALL.

25 [1] Menon N.K., Robbins J., Peck H.D. Jr., Chatelus C.Y., Choi E.-S., Przybyla A.E. J. Bacteriol. 172:1969-1977(1990).

[2] Volbeda A., Charon M.-H., Piras C., Hatchikian E.C., Frey M., Fontecilla-Camps J.C. Nature 373:580-587(1995).

[3] Eidsness M.K., Scott R.A., Prickrill B., der Vartanian D.V., LeGall J., Moura I., Moura J.J.G., Peck H.D. Jr. Proc. Natl. Acad. Sci. U.S.A. 86:147-151(1989).
30

[4] Tran-Betcke A., Warnecke U., Boecker C., Zaborosch C., Friedrich B. J. Bacteriol. 172:2920-2929(1990).

791

976. NADH-Ubiquinone oxidoreductase (complex I), chain 5 C-terminus (oxidored_q1_C)

This sub-family represents a carboxyl terminal extension of oxidored_q1. Only NADH-Ubiquinone chain 5 from chloroplasts are in this family. This sub-family is part of complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane.

Number of members: 572

[1] Walker JE; Medline: 93110040 "The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains." Q Rev Biophys 1992;25:253-324.

977. NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminus (oxidored_q1_N)

This sub-family represents an amino terminal extension of oxidored_q1. Only NADH-Ubiquinone chain 5 and eubacterial chain L are in this family. This sub-family is part of complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane.

Number of members: 546

[1] Walker JE; Medline: 93110040 "The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains." Q Rev Biophys 1992;25:253-324.

978. oxidored_q2. NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 4L (EC 1.6.5.3). ND4L OR NAD4L. Arabidopsis thaliana (Mouse-ear cress). Mitochondrion. OC Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Rosidae; eurosids II; Brassicales; Brassicaceae; Arabidopsis. CATALYTIC ACTIVITY: NADH + UBIQUINONE = NAD(+) + UBIQUINOL.

[1] SEQUENCE FROM N.A. MEDLINE; 93156682. Brandt P., Sunkel S., Unseld M., Brennicke A., Knoop V.; "The nad4L gene is encoded between exon c of nad5 and orf25 in the Arabidopsis mitochondrial genome."; Mol. Gen. Genet. 236:33-38(1992).

[2] SEQUENCE FROM N.A. STRAIN=CV. COLUMBIA; MEDLINE; 97141919 Unseld M., Marienfeld J.R., Brandt P., Brennicke A.; "The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides."; Nat. Genet. 15:57-61(1997).

- 5 979. oxidored_q4. Protein name NADH-PLASTOQUINONE OXIDOREDUCTASE CHAIN 3, CHLOROPLAST. Synonym(s) EC 1.6.5.3. Gene name(s) NDHC OR NDH3 From *Zea mays* (Maize) Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; *Zea*.

CATALYTIC ACTIVITY: NADH + PLASTOQUINONE = NAD(+) +

- 10 PLASTOQUINOL.

SIMILARITY: BELONGS TO THE COMPLEX I SUBUNIT 3 FAMILY.

[1] SEQUENCE FROM N.A. MEDLINE; 89281491. Steinmueller K., Ley A.C., Steinmetz A.A., Sayre R.T., Bogorad L.; "Characterization of the *ndhC-psbG-ORF157/159* operon of maize plastid DNA and of the cyanobacterium *Synechocystis* sp. PCC6803."; Mol. Gen. Genet. 216:60-69(1989).

[2] SEQUENCE FROM N.A. MEDLINE; 95395841. Maier R.M., Neckermann K., Igloi G.L., Koessel H.; "Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing."; J. Mol. Biol. 251:614-628(1995).

980. PAC: PAC motif

PAC motif occurs C-terminal to a subset of all known PAS motifs. It is proposed to contribute to the PAS domain fold [3]. Number of members: 181

[1] Medline: 97446881 PAS domain S-boxes in archaea, bacteria and sensors for oxygen and redox. Zhulin IB, Taylor BL, Dixon R; Trends Biochem Sci 1997;22:331-333.

[2] Medline: 95275818. 1.4 A structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore. Borgstahl GE, Williams DR, Getzoff ED; Biochemistry 1995;34:6278-6287.

[3] Medline: 98044337. PAS: a multifunctional domain family comes to light. Ponting CP, Aravind L; Curr Biol 1997;7:674-677.

793

981. PARP: Poly(ADP-ribose) polymerase catalytic region.

Poly(ADP-ribose) polymerase catalyses the covalent attachment of ADP-ribose units from NAD⁺ to itself and to a limited number of other DNA binding proteins, which decreases their affinity for DNA. Poly(ADP-ribose) polymerase is a regulatory component induced by DNA damage.

The carboxyl-terminal region is the most highly conserved region of the protein. Experiments have shown that a carboxyl 40 kDa fragment is still catalytically active [2]. Number of members: 19

[1] Medline: 96353841 Structure of the catalytic fragment of poly(AD-ribose) polymerase from chicken. Ruf A, Mennissier de Murcia J, de Murcia G, Schulz GE; Proc Natl Acad Sci U S A 1996;93:7481-7485.

[2] Medline: 93293867 The carboxyl-terminal domain of human poly(ADP-ribose) polymerase. Overproduction in Escherichia coli, large scale purification, and characterization. Simonin F, Hofferer L, Panzeter PL, Muller S, de Murcia G, Althaus FR; J Biol Chem 1993;268:13454-13461.

982. PC_rep: Proteasome/cyclosome repeat

[1] Medline: 97348748 A repetitive sequence in subunits of the 26S proteasome and 20S cyclosome (anaphase-promoting complex). Lupas A, Baumeister W, Hofmann K; Trends Biochem Sci 1997;22:195-196.

Number of members: 112

983. Peptidase_M1: Peptidase family M1

Members of this family are aminopeptidases. The members differ widely in specificity, hydrolysing acidic, basic or neutral N-terminal residues. This family includes leukotriene-A4 hydrolase Swiss:P09960, this enzyme also has an aminopeptidase activity [1]. Number of members: 72

[1] Medline: 95405261 Evolutionary families of metallopeptidases. Rawlings ND, Barrett AJ; Meth Enzymol 1995;248:183-228.

984. Neutral zinc metallopeptidases, zinc-binding region signature (Peptidase_M8)

PROSITE cross-reference(s) PS00142; ZINC_PROTEASE

The majority of zinc-dependent metallopeptidases (with the notable exception of the carboxypeptidases) share a common pattern of primary structure [1,2,3] in the part of their sequence involved in the binding of zinc, and can be grouped together as a superfamily, known as the metzincins, on the basis of this sequence similarity. They can be classified into a number of distinct families [4,E1] which are listed below along with the proteases which are currently known to belong to these families.

Family M1

- Bacterial aminopeptidase N (EC 3.4.11.2) (gene pepN).
- Mammalian aminopeptidase N (EC 3.4.11.2).
- Mammalian glutamyl aminopeptidase (EC 3.4.11.7) (aminopeptidase A). It may play a role in regulating growth and differentiation of early B-lineage cells.
- Yeast aminopeptidase yscII (gene APE2).
- Yeast alanine/arginine aminopeptidase (gene AAP1).
- Yeast hypothetical protein YIL137c.
- Leukotriene A-4 hydrolase (EC 3.3.2.6). This enzyme is responsible for the hydrolysis of an epoxide moiety of LTA-4 to form LTB-4; it has been shown that it binds zinc and is capable of peptidase activity.

Family M2

- Angiotensin-converting enzyme (EC 3.4.15.1) (dipeptidyl carboxypeptidase I) (ACE) the enzyme responsible for hydrolyzing angiotensin I to angiotensin II. There are two forms of ACE: a testis-specific isozyme and a somatic isozyme which has two active centers.

Family M3

- Thimet oligopeptidase (EC 3.4.24.15), a mammalian enzyme involved in the cytoplasmic degradation of small peptides.
- Neurolysin (EC 3.4.24.16) (also known as mitochondrial oligopeptidase M or microsomal endopeptidase).
- Mitochondrial intermediate peptidase precursor (EC 3.4.24.59) (MIP). It is involved the second stage of processing of some proteins imported in the mitochondrion.
- Yeast saccharolysin (EC 3.4.24.37) (proteinase yscD).

795

- *Escherichia coli* and related bacteria dipeptidyl carboxypeptidase (EC 3.4.15.5) (gene dcp).
- *Escherichia coli* and related bacteria oligopeptidase A (EC 3.4.24.70) (gene opdA or prlC).
- Yeast hypothetical protein YKL134c.

5 Family M4

- Thermostable thermolysins (EC 3.4.24.27), and related thermolabile neutral proteases (bacillolysins) (EC 3.4.24.28) from various species of *Bacillus*.
- Pseudolysin (EC 3.4.24.26) from *Pseudomonas aeruginosa* (gene lasB).
- Extracellular elastase from *Staphylococcus epidermidis*.
- Extracellular protease prt1 from *Erwinia carotovora*.
- Extracellular minor protease smp from *Serratia marcescens*.
- Vibriolysin (EC 3.4.24.25) from various species of *Vibrio*.
- Protease prtA from *Listeria monocytogenes*.
- Extracellular proteinase proA from *Legionella pneumophila*.

15 Family M5

- Mycolysin (EC 3.4.24.31) from *Streptomyces cacaoi*.

20 Family M6

- Immune inhibitor A from *Bacillus thuringiensis* (gene ina). Ina degrades two classes of insect antibacterial proteins, attacins and cecropins.

25 Family M7

- *Streptomyces* extracellular small neutral proteases

Family M8

- Leishmanolysin (EC 3.4.24.36) (surface glycoprotein gp63), a cell surface protease from various species of *Leishmania*.

30 Family M9

- Microbial collagenase (EC 3.4.24.3) from *Clostridium perfringens* and *Vibrio alginolyticus*.

Family M10A

- Serralysin (EC 3.4.24.40), an extracellular metalloprotease from *Serratia*.
- Alkaline metalloproteinase from *Pseudomonas aeruginosa* (gene *aprA*).
- Secreted proteases A, B, C and G from *Erwinia chrysanthemi*.
- 5 - Yeast hypothetical protein YIL108w.

Family M10B

- Mammalian extracellular matrix metalloproteinases (known as matrixins) [5]: MMP-1 (EC 3.4.24.7) (interstitial collagenase), MMP-2 (EC 3.4.24.24) (72 Kd gelatinase), MMP-9 (EC 10 3.4.24.35) (92 Kd gelatinase), MMP-7 (EC 3.4.24.23) (matrylisin), MMP-8 (EC 3.4.24.34) (neutrophil collagenase), MMP-3 (EC 3.4.24.17) (stromelysin-1), MMP-10 (EC 3.4.24.22) (stromelysin-2), and MMP-11 (stromelysin-3), MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).
- Sea urchin hatching enzyme (envelysin) (EC 3.4.24.12). A protease that allows the embryo to digest the protective envelope derived from the egg extracellular matrix.
- Soybean metalloendoproteinase 1.

Family M11

- *Chlamydomonas reinhardtii* gamete lytic enzyme (GLE).

Family M12A

- Astacin (EC 3.4.24.21), a crayfish endoprotease.
- Meprin A (EC 3.4.24.18), a mammalian kidney and intestinal brush border metalloendopeptidase.
- 25 - Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity. The *Drosophila* homolog of BMP-1 is the dorsal-ventral patterning protein *tolloid*.
- Blastula protease 10 (BP10) from *Paracentrotus lividus* and the related protein SpAN from *Strongylocentrotus purpuratus*.
- 30 - *Caenorhabditis elegans* protein *toh-2*.
- *Caenorhabditis elegans* hypothetical protein F42A10.8.
- Choriolysins L and H (EC 3.4.24.67) (also known as embryonic hatching proteins LCE and HCE) from the fish *Oryzias latipes*. These proteases participate in the breakdown

of the egg envelope, which is derived from the egg extracellular matrix, at the time of hatching.

Family M12B

- 5 - Snake venom metalloproteinases [6]. This subfamily mostly groups proteases that act in hemorrhage. Examples are: adamalysin II (EC 3.4.24.46), atrolysin C/D (EC 3.4.24.42), atrolysin E (EC 3.4.24.44), fibrolase (EC 3.4.24.72), trimereylisin I (EC 3.4.25.52) and II (EC 3.4.25.53).
- 10 - Mouse cell surface antigen MS2.

Family M13

- 15 - Mammalian neprilysin (EC 3.4.24.11) (neutral endopeptidase) (NEP).
- 20 - Endothelin-converting enzyme 1 (EC 3.4.24.71) (ECE-1), which process the precursor of endothelin to release the active peptide.
- 25 - Kell blood group glycoprotein, a major antigenic protein of erythrocytes. The Kell protein is very probably a zinc endopeptidase.
- 30 - Peptidase O from *Lactococcus lactis* (gene pepO).

Family M27

- 35 - Clostridial neurotoxins, including tetanus toxin (TeTx) and the various botulinum toxins (BoNT). These toxins are zinc proteases that block neurotransmitter release by proteolytic cleavage of synaptic proteins such as synaptobrevins, syntaxin and SNAP-25 [7,8].

Family M30

- 40 - *Staphylococcus hyicus* neutral metalloprotease.

Family M32

- 45 - Thermostable carboxypeptidase 1 (EC 3.4.17.19) (carboxypeptidase Taq), an enzyme from *Thermus aquaticus* which is most active at high temperature.

Family M34

- Lethal factor (LF) from *Bacillus anthracis*, one of the three proteins composing the anthrax toxin.

Family M35

- 5 - Deuterolysin (EC 3.4.24.39) from *Penicillium citrinum* and related proteases from various species of *Aspergillus*.

Family M36

- Extracellular elastinolytic metalloproteinases from *Aspergillus*.

10

From the tertiary structure of thermolysin, the position of the residues acting as zinc ligands and those involved in the catalytic activity are known. Two of the zinc ligands are histidines which are very close together in the sequence; C-terminal to the first histidine is a glutamic acid residue which acts as a nucleophile and promotes the attack of a water molecule on the carbonyl carbon of the substrate. A signature pattern which includes the two histidine and the glutamic acid residues is sufficient to detect this superfamily of proteins.

15

Consensus pattern[GSTALIVN]-x(2)-H-E-[LIVMFYW]-{DEHRKP}-H-x-
20 [LIVMFYWGSPQ]

[The two H's are zinc ligands] [E is the active site residue]

Sequences known to belong to this class detected by the patternALL, except for members of families M5, M7 and M11.

Other sequence(s) detected in SWISS-PROT57; including *Neurospora crassa*
25 conidiation-specific protein 13 which could be a zinc-protease.

[1]Jongeneel C.V., Bouvier J., Bairoch A. FEBS Lett. 242:211-214(1989).

[2]Murphy G.J.P., Murphy G., Reynolds J.J. FEBS Lett. 289:4-7(1991).

[3]Bode W., Grams F., Reinemer P., Gomis-Rueth F.-X., Baumann U., McKay D.B., Stoecker W. Zoology 99:237-246(1996).

30 [4]Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

[5]Woessner J. Jr. FASEB J. 5:2145-2154(1991).

[6]Hite L.A., Fox J.W., Bjarnason J.B. Biol. Chem. Hoppe-Seyler 373:381-385(1992).

[7]Montecucco C., Schiavo G. Trends Biochem. Sci. 18:324-327(1993).

[8]Niemann H., Blasi J., Jahn R. Trends Cell Biol. 4:179-185(1994).

985. PHO4: Phosphate transporter family

This family includes PHO-4 from *Neurospora crassa* which is a Na(+)-phosphate symporter [1]. This family also contains the leukemia virus receptor Swiss:Q08344. Number of members: 41

[1] Medline: 95249577 Repressible cation-phosphate symporters in *Neurospora crassa*. Versaw WK, Metzenberg RL; Proc Natl Acad Sci U S A 1995;92:3884-3887.

986. Photosynthetic reaction center proteins signature (photoRC)

PROSITE cross-reference(s): PS00244; REACTION_CENTER

In the photosynthetic reaction center of purple bacteria, two homologous integral membrane proteins, L(ight) and M(edium), are known to be essential to the light-mediated water-splitting process. In the photosystem II of eukaryotic chloroplasts two related proteins are involved: the D1 (psbA) and D2 proteins (psbD). These four types of protein probably evolved from a common ancestor [see 1,2 for recent reviews].

A signature pattern was developed which include two conserved histidine residues. In L and M chains, the first histidine is a ligand of the magnesium ion of the special pair bacteriochlorophyll, the second is a ligand of a ferrous non-heme iron atom. In photosystem II these two histidines are thought to play a similar role.

Consensus pattern[NQH]-x(4)-P-x-H-x(2)-[SAG]-x(11)-[SAGC]-x-H-[SAG](2)
[The first H is a magnesium ligand] [The second H is a iron ligand]
Sequences known to belong to this class detected by the patternALL, except for broad bean psbA which has Gln instead of the second His.

[1]Michel H., Deisenhofer J. Biochemistry 27:1-7(1988).
[2]Barber J. Trends Biochem. Sci. 12:321-326(1987).

987. phytochrome: Phytochrome region

800

This family contains a region specific to phytochrome proteins. Number of members:
145

988. PI3K_C2: C2 domain

- 5 Phosphoinositide 3-kinase region postulated to contain a C2 domain. Outlier of C2 family.
Number of members: 39

[1] Medline: 97388296 Using structure to define the function of phosphoinositide 3-kinase family members. Domin J, Waterfield MD; FEBS Lett 1997;410:91-95.

- 10 [2] Medline: 97398940 Phosphoinositide 3-kinases: a conserved family of signal transducers. Vanhaesebroeck B, Leever SJ, Panayotou G, Waterfield MD; Trends Biochem Sci 1997;22:267-272.

989. PI3Ka: Phosphoinositide 3-kinase family, accessory domain (PIK domain)

- 15 PIK domain is conserved in all PI3 and PI4-kinases. Its role is unclear but it has been suggested [2] to be involved in substrate presentation.
Number of members: 47

[1] Medline: 97388296 Using structure to define the function of phosphoinositide 3-kinase family members. Domin J, Waterfield MD; FEBS Lett 1997;410:91-95.

- 20 [2] Medline: 94069320 Phosphatidylinositol 4-kinase: gene structure and requirement for yeast cell viability. Flanagan CA, Schnieders EA, Emerick AW, Kunisawa R, Admon A, Thorner J; Science 1993;262:1444-1448.

25 990. P-II protein signatures

PROSITE cross-reference(s): PS00496; PII_GLNB_UMP, PS00638; PII_GLNB_CTER

The P-II protein (gene glnB) is a bacterial protein important for the control of glutamine synthetase [1,2,3]. In nitrogen-limiting conditions, when the ratio of glutamine to 2-ketoglutarate decreases, P-II is uridylylated on a tyrosine residue to form P-II-UMP. P-II-UMP allows the deadenylation of glutamine synthetase (GS), thus activating the enzyme. Conversely, in nitrogen excess, P-II-UMP is deuridylylated and then promotes the adenylation of GS. P-II also indirectly controls the transcription of the GS gene (glnA) by preventing NR-

30

II (ntrB) to phosphorylate NR-I (ntrC) which is the transcriptional activator of glnA. Once P-II is uridylylated, these events are reversed.

P-II is a protein of about 110 amino acid residues extremely well conserved. The tyrosine which is uridylylated is located in the central part of the protein.

In cyanobacteria, P-II seems to be phosphorylated on a serine residue rather than being uridylylated.

In methanogenic archaeobacteria, the nitrogenase iron protein gene (nifH) is followed by two open reading frames highly similar to the eubacterial P-II protein [4]. These proteins could be involved in the regulation of nitrogen fixation.

In the red alga, *Porphyra purpurea*, there is a glnB homolog encoded in the chloroplast genome.

Other proteins highly similar to glnB are:

- *Bacillus subtilis* protein nrgB [5].
- *Escherichia coli* hypothetical protein ybaI [6].

Two signature patterns were developed for P-II protein. The first one is a conserved stretch (in eubacteria) of six residues which contains the uridylylated tyrosine, the other is derived from a conserved region in the C-terminal part of the P-II protein.

Consensus pattern Y-[KR]-G-[AS]-[AE]-Y [The second Y is uridylylated]

Sequences known to belong to this class detected by the pattern ALL glnB's from eubacteria.

Consensus pattern [ST]-x(3)-G-[DY]-G-[KR]-[IV]-[FW]-[LIVM]-x(2)-[LIVM]

[1] Magasanik B. *Biochimie* 71:1005-1012(1989).

[2] Holtel A., Merrick M. *Mol. Gen. Genet.* 215:134-138(1988).

[3]Cheah E., Carr P.D., Suffolk P.M., Vasuvedan S.G., Dixon N.E., Ollis D.L. Structure 2:981-990(1994).

[4]Sibold L., Henriquet M., Possot O., Aubert J.-P. Res. Microbiol. 142:5-12(1991).

[5]Wray L.V. Jr., Atkinson M.R., Fisher S.H. J. Bacteriol. 176:108-114(1994).

5 [6]Allikmets R., Gerrard B.C., Court D., Dean M.C. Gene 136:231-236(1993).

991. PIP5K: Phosphatidylinositol-4-phosphate 5-Kinase

This family contains a region from the common kinase core found in the type I phosphatidylinositol-4-phosphate 5-kinase (PIP5K) family as described in [1]. The family
10 consists of various type I, II and III PIP5K enzymes. PIP5K catalyses the formation of phosphoinositol-4,5-bisphosphate via the phosphorylation of phosphatidylinositol-4-phosphate a precursor in the phosphoinositide signaling pathway. Number of members: 33

[1] Medline: 98204859. Type I phosphatidylinositol-4-phosphate 5-kinases. Cloning of the
15 third isoform and deletion/substitution analysis of members of this novel lipid kinase family. Ishihara H, Shibasaki Y, Kizuki N, Wada T, Yazaki Y, Asano T, Oka Y; J Biol Chem 1998;273:8741-8748.

[2] Medline: 97115834 Type I phosphatidylinositol-4-phosphate 5-kinases are distinct
20 members of this novel lipid kinase family. Loijens JC, Anderson RA; J Biol Chem 1996 20;271:32937-32943.

992. PolyA_pol: Poly A polymerase family

This family includes nucleic acid independent RNA polymerases, such as Poly(A) polymerase, which adds the poly (A) tail to mRNA EC:2.7.7.19. This family also includes the
25 tRNA nucleotidyltransferase that adds the CCA to the 3' of the tRNA EC:2.7.7.25. Number of members: 31

[1] Medline: 93066242 Identification of the gene for an Escherichia coli poly(A) polymerase. Cao GJ, Sarkar N; Proc Natl Acad Sci U S A 1992;89:10380-10384.

30 993. Photosystem I psaA and psaB proteins signature (psaA_psaB)
PROSITE cross-reference(s)PS00419; PHOTOSYSTEM_I_PSAAB

Photosystem I (PSI) [1] is an integral membrane protein complex that uses light energy to mediate electron transfer from plastocyanin to ferredoxin. PSI is found in the chloroplast of plants and cyanobacteria. The electron transfer components of the reaction center of PSI are a primary electron donor P-700 (chlorophyll dimer) and five electron acceptors: A0 (chlorophyll), A1 (a phylloquinone) and three 4Fe-4S iron-sulfur centers: Fx, Fa, and Fb.

PsaA and psaB, two closely related proteins, are involved in the binding of P700, A0, A1, and Fx. psaA and psaB are both integral membrane proteins of 730 to 750 amino acids that seem to contain 11 transmembrane segments. The Fx 4Fe-4S iron-sulfur center is bound by four cysteines; two of these cysteines are provided by the psaA protein and the two others by psaB. The two cysteines in both proteins are proximal and located in a loop between the ninth and tenth transmembrane segments. A leucine zipper motif seems to be present [2] downstream of the cysteines and could contribute to dimerization of psaA/psaB.

The signature pattern for these proteins is based on the perfectly conserved region that includes the two iron-sulfur binding cysteines.

Consensus pattern C-D-G-P-G-R-G-G-T-C [The two C's bind the iron-sulfur center]

[1] Golbeck J.H. Biochim. Biophys. Acta 895:167-204(1987).

[2] Webber A.N., Malkin R. FEBS Lett. 264:1-14(1990).

994. PSBH: Photosystem II 10 kDa phosphoprotein

This protein is phosphorylated in a light dependent reaction.

Number of members: 20

995. PsbJ

This family consists of the photosystem II reaction center protein PsbJ from plants and Cyanobacteria. In *Synechocystis* sp. PCC 6803 PsbJ regulates the number of photosystem II centers in thylakoid membranes, it is a predicted 4kDa protein with one membrane spanning domain [1]. Number of members: 20

[1] Medline: 93131892. Genetic and immunological analyses of the cyanobacterium *Synechocystis* sp. PCC 6803 show that the protein encoded by the psbJ gene regulates the

number of photosystem II centers in thylakoid membranes. Lind LK, Shukla VK, Nyhus KJ, Pakrasi HB; J Biol Chem 1993;268:1575-1579.

996. PSBT: Photosystem II reaction centre T protein

- 5 The exact function of this protein is unknown. It probably consists of a single transmembrane spanning helix. The Swiss:P37256 protein, appears to be (i) a novel photosystem II subunit and (ii) required for maintaining optimal photosystem II activity under adverse growth conditions [1]. Number of members: 17

- 10 [1] Medline: 94298765. The chloroplast ycf8 open reading frame encodes a photosystem II polypeptide which maintains photosynthetic activity under adverse growth conditions. Monod C, Takahashi Y, Goldschmidt-Clermont M, Rochaix JD; EMBO J 1994;13:2747-2754.

- 5 997. PSI_8. PHOTOSYSTEM I REACTION CENTRE SUBUNIT VIII. Synonym(s)PSI-I. Gene name(s)PSAI. From Hordeum vulgare (Barley). Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; Hordeum.
MAY HELP IN THE ORGANIZATION OF THE PSAL SUBUNIT. BELONGS TO THE
20 PSAL FAMILY.

[1] SEQUENCE FROM N.A. MEDLINE; 90036933. Scheller H.V., Okkels J.S., Hoej P.B., Svendsen I., Roepstorff P., Moeller B.L.; "The primary structure of a 4.0-kDa photosystem I polypeptide encoded by the chloroplast psal gene."; J. Biol. Chem. 264:18402-18406(1989).

25

998. PSI_PsaJ: Photosystem I reaction centre subunit IX / PsaJ

This family consists of the photosystem I reaction centre subunit IX or PsaJ from various organisms including Synechocystis sp. (strain pcc 6803), Pinus thunbergii (green pine) and Zea mays (maize). PsaJ Swiss:P19443 is a small 4.4kDa, chloroplastal encoded, hydrophobic

30

subunit of the photosystem I reaction complex its function is not yet fully understood [1].

PsaJ can be cross-linked to PsaF Swiss:P12356 and has a single predicted transmembrane domain it has a proposed role in maintaing PsaF in the correct orientation to allow for fast electron transfer from soluble donor proteins to P700+ [1]. Number of members: 18

[1] Medline: 99238330. A large fraction of PsaF is nonfunctional in photosystem I complexes lacking the PsaJ subunit. Fischer N, Boudreau E, Hippler M, Drepper F, Haehnel W, Rochaix JD; Biochemistry 1999;38:5546-5552.

- 5 [2] Medline: 93252282. Genes encoding eleven subunits of photosystem I from the thermophilic cyanobacterium *Synechococcus* sp. Muhlenhoff U, Haehnel W, Witt H, Herrmann RG; Gene 1993;127:71-78.

999. PSII. Protein namePHOTOSYSTEM II P680 CHLOROPHYLL A APOPROTEIN.

- 10 Synonym(s)CP-47 PROTEIN. Gene name(s)PSBB. From *Hordeum vulgare* (Barley), Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; Hordeum.

5 FUNCTION: THIS PROTEIN CONJUGATES WITH CHLOROPHYLL & CATALYZES THE PRIMARY LIGHT-INDUCED PHOTOCHEMICAL PROCESSES OF PHOTOSYSTEM II. SUBCELLULAR LOCATION: CHLOROPLAST THYLAKOID MEMBRANE. SIMILARITY: BELONGS TO THE PSBB / PSBC FAMILY.

20 [1] SEQUENCE FROM N.A. STRAIN=CV. SABARLIS; MEDLINE; 89240047. Andreeva A.V., Buryakova A.A., Reverdatto S.V., Chakhmakhcheva O.G., Efimov V.A.; "Nucleotide sequence of the 5.2 kbp barley chloroplast DNA fragment, containing psbB-psbH-petB-petD gene cluster."; Nucleic Acids Res. 17:2859-2860(1989).

[2] SEQUENCE FROM N.A. STRAIN=CV. SABARLIS; MEDLINE; 92207253. Efimov V.A., Andreeva A.V., Reverdatto S.V., Chakhmakhcheva O.G.; "Photosystem II of rye. Nucleotide sequence of the psbB, psbC, psbE, psbF, psbH genes of rye and chloroplast DNA regions adjacent to them."; Bioorg. Khim. 17:1369-1385(1991).

25 [3] SEQUENCE OF 411-420. Hinz U.G.; "Isolation of the photosystem II reaction center complex from barley. Characterization by circular dichroism spectroscopy and amino acid sequencing."; Carlsberg Res. Commun. 50:285-298(1985).

- 30 1000. QRPTase. Quinolate phosphoribosyl transferase.
Quinolate phosphoribosyl transferase (QRPTase) or nicotinate-nucleotide pyrophosphorylase EC:2.4.2.19 is involved in the de novo synthesis of NAD in both prokaryotes and eukaryotes. It catalyses the reaction of quinolinic acid with 5-

806

phosphoribosyl-1-pyrophosphate (PRPP) in the presence of Mg^{2+} to give rise to nicotinic acid mononucleotide (NaMN), pyrophosphate and carbon dioxide [1,2]. Number of members: 26.

5 [1]Medline: 97169443. A new function for a common fold: the crystal structure of quinolinic acid phosphoribosyltransferase. Eads JC, Ozturk D, Wexler TB, Grubmeyer C, Sacchettini JC; Structure 1997;5:47-58.

[2]Medline: 96139309. The sequencing expression, purification, and steady-state kinetic analysis of quinolinate phosphoribosyl transferase from Escherichia coli. Bhatia R, Calvo
10 KC; Arch Biochem Biophys 1996;325:270-278.

1001. R3H domain

The name of the R3H domain comes from the characteristic spacing of the most conserved arginine and histidine residues. The function of the domain is predicted to be binding
15 ssDNA. Number of members: 28

[1]Medline: 99003905 The R3H motif: a domain that binds single-stranded nucleic acids. Grishin NV; Trends Biochem Sci 1998;23:329-330.

20 1002. recF protein signatures (RecF)

The prokaryotic protein recF [1,2] is a single-stranded DNA-binding protein which also probably binds ATP. RecF is involved in DNA metabolism; it is required for recombinational DNA repair and for induction of the SOS response. RecF is a protein of about 350 to 370
25 amino acid residues; there is a conserved ATP-binding site motif 'A' (P-loop) in the N-terminal section of the protein as well as two other conserved regions, one located in the central section, and the other in the C-terminal section. Signature patterns were derived from these two regions.

30 Consensus pattern [LIVM]-x(4)-[LIF]-x(6)-[LIF]-[LVF]-x-[GE]-[GSTAD]-[PA]- x(2)-R-R-x-[FYW]-[LIVMF]-D Sequences known to belong to this class detected by the pattern ALL.

807

Consensus pattern[LIVMFY](2)-x-D-x(2,3)-[SA]-[EH]-L-D-x(2)-[KRH]-x(3)-L Sequences known to belong to this class detected by the patternALL, except for T. palidum recF.

[1] Sandler S.J., Chackerian B., Li J.T., Clark A.J. Nucleic Acids Res. 20:839-845(1992).

5 [2] Alonso J.C., Fisher L.M.; Mol. Gen. Genet. 246:680-686(1995).

1003. RibD C-terminal domain (RibD_C)

The function of this domain is not known, but it is thought to be involved in riboflavin biosynthesis. This domain is found in the C terminus of RibD/RibG Swiss:P25539, in combination with dCMP_cyt_deam, as well as in isolation in some archaeobacterial proteins Swiss:P95872.

Number of members: 21

1004. Ribosomal protein L16 signatures (Ribosomal_L16)

Ribosomal protein L16 is one of the proteins from the large ribosomal subunit. In Escherichia coli, L16 is known to bind directly the 23S rRNA and to be located at the A site of the peptidyltransferase center. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:

- Eubacterial L16.
- Algal and plant chloroplast L16.
- Cyanelle L16.
- Plant mitochondrial L16.

L16 is a protein of 133 to 185 amino-acid residues. As signature patterns, we selected two conserved regions in the central section of these proteins.

Consensus pattern [KR](2)-x-[GSAC]-[KRQVA]-[LIVM]-W-[LIVM]-[KR]-[LIVM]-[LFY]-[AP] Sequences known to belong to this class detected by the pattern ALL.

Consensus patternR-M-G-x-[GR]-K-G-x(4)-[FWKR] Sequences known to belong to this class detected by the patternALL.

[1] Otaka E., Hashimoto T., Mizuta K., Suzuki K. Protein Seq. Data Anal. 5:301-313(1993).

1005. Ribosomal protein L32c signature (Ribosomal_L32E)

5 A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Mammalian L32 [1].
- Drosophila RP49 [2].
- Trichoderma harzianum L32 [3].

10 - Yeast L32c (YBL092w).

- Archaeobacterial L32c [4].

These proteins have 135 to 240 amino-acid residues. As a signature pattern, a stretch of about 20 residues located in the N-terminal part of these proteins was selected.

15 Consensus pattern F-x-R-x(4)-[KR]-x(2)-[KR]-[LIVMF]-x(3,5)-W-R-[KR]-x(2)-G Sequences known to belong to this class detected by the pattern ALL.

[1] Jacks C.M., Powasir C.B., Hackett P.B. Gene 74:565-570(1988).

[2] Aguade M. Mol. Biol. Evol. 5:433-441(1988).

20 [3] Lora J.M., Garcia L., Benitez T., Llobell A., Pintor-Toro J.A. Nucleic Acids Res. 21:3319-3319(1993).

[4] Arndt E., Scholzen T., Kroemer W., Hatakeyama T., Kimura M. Biochimie 73:657-668(1991).

25 1006. (Ribosomal_S3) Ribosomal protein S3 signature

PROSITE: PDOC00474. PROSITE cross-reference(s) PS00548; RIBOSOMAL_S3

Ribosomal protein S3 is one of the proteins from the small ribosomal subunit.

In Escherichia coli, S3 is known to be involved in the binding of initiator Met-tRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1],

30 groups:

- Eubacterial S3.
- Algal and plant chloroplast S3.
- Cyanelle S3.

-Archaeobacterial S3.

-Plant mitochondrial S3.

-Vertebrate S3.

-Insect S3.

5 -Caenorhabditis elegans S3 (C23G10.3).

-Yeast S3 (Rp13).

S3 is a protein of 209 to 559 amino-acid residues. A conserved region located in the C-terminal section was selected as a signature pattern.

10 Consensus pattern[GSTA]-[KR]-x(6)-G-x-[LIVMT]-x(2)-[NQSCH]-x(1,3)-[LIVFCA]-x(3)-[LIV]-[DENQ]-x(7)-[LMT]-x(2)-G-x(2)-[GS]. Sequences known to belong to this class detected by the patternALL, except for some mitochondrial S3.

[1]Otaka E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).

15 1007. RimM - RimM

The RimM protein is essential for efficient processing of 16S rRNA [1]. The RimM protein was shown to have affinity for free ribosomal 30S subunits but not for 30S subunits in the 70S ribosomes [1]. Number of members: 14.

20 [1]Medline: 98083058. RimM and RbfA are essential for efficient processing of 16S rRNA in Escherichia coli. Bylund GO, Wipemo LC, Lundberg LA, Wikstrom PM; J Bacteriol 1998;180:73-82.

25 1008. RNA_pol_A - RNA polymerase alpha subunit

-!- RNA polymerases catalyse the DNA dependent polymerisation of RNA. Prokaryotes contain a single RNA polymerase compared to three in eukaryotes (not including mitochondrial and chloroplast polymerases).

30 -!- Members of this family include: A subunit from eukaryotes, gamma subunit from cyanobacteria, beta' subunit from eubacteria, A' subunit from archaeobacteria, B" from chloroplasts. Number of members: 139.

[1]Medline: 97066998. Structural modules of the large subunits of RNA polymerase. Introducing archaebacterial and chloroplast split sites in the beta and beta' subunits of Escherichia coli RNA polymerase. Severinov K, Mustaev A, Kukarin A, Muzzin O, Bass I, Darst SA, Goldfarb A; J Biol Chem 1996;271:27969-27974.

5

1009. RuBisCO_large - Ribulose biphosphate carboxylase large chain active site
PROSITE: PDOC00142; PROSITE cross-reference(s) PS00157; RUBISCO_LARGE

Ribulose biphosphate carboxylase (EC 4.1.1.39) (RuBisCO) [1,2] catalyzes the initial step in Calvin's reductive pentose phosphate cycle in plants as well as purple and green bacteria. It consists of a large catalytic unit and a small subunit of undetermined function. In plants, the large subunit is coded by the chloroplastic genome while the small subunit is encoded in the nuclear genome. Molecular activation of RuBisCO by CO₂ involves the formation of a carbamate with the epsilon-amino group of a conserved lysine residue. This carbamate is stabilized by a magnesium ion. One of the ligands of the magnesium ion is an aspartic acid residue close to the active site lysine [3]. A pattern was developed which includes both the active site residue and the metal ligand, and which is specific to RuBisCO large chains.

15

20

Consensus pattern G-x-[DN]-F-x-K-x-D-E [K is the active site residue] [The second D is a magnesium ligand]. Sequences known to belong to this class detected by the pattern ALL, except for Cheilopleuria bicuspidis RuBisCO.

25

[1]Miziorko H.M., Lorimer G.H. Annu. Rev. Biochem. 52:507-535(1983).

[2]Akazawa T., Takabe T., Kobayashi H. Trends Biochem. Sci. 9:380-383(1984).

[3]Andersson L., Knight S., Schneider G., Lindqvist Y., Lundqvist T., Branden C.-I., Lorimer G.H. Nature 337:229-234(1989).

30

1010. Rve - Integrase core domain

Integrase mediates integration of a DNA copy of the viral genome into the host chromosome.

Integrase is composed of three domains. The amino-terminal domain is a zinc binding domain Integrase_Zn. This domain is the central catalytic domain. The carboxyl terminal domain that is a non-specific DNA binding domain integrase. The catalytic domain acts as an endonuclease when two nucleotides are removed from the 3' ends of the blunt-ended viral

DNA made by reverse transcription. This domain also catalyses the DNA strand transfer reaction of the 3' ends of the viral DNA to the 5' ends of the integration site [1]. Number of members: 694.

- 5 [1]Medline: 95099322. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR; Science 1994;266:1981-1986.

1011. (SBP_bac_3) Bacterial extracellular solute-binding proteins, family 3 signature

10 PROSITE: PDOC00798. PROSITE cross-reference(s) PS01039; SBP_BACTERIAL_3

Bacterial high affinity transport systems are involved in active transport of solutes across the cytoplasmic membrane. The protein components of these traffic systems include one or two transmembrane protein components, one or two membrane-associated ATP-binding proteins (ABC transporters; see <PDOC00185>) and a high affinity periplasmic solute-binding protein. The later are thought to bind the substrate in the vicinity of the inner membrane, and to transfer it to a complex of inner membrane proteins for concentration into the cytoplasm.

In gram-positive bacteria which are surrounded by a single membrane and have therefore no periplasmic region the equivalent proteins are bound to the membrane via an N-terminal lipid anchor. These homolog proteins do not play an integral role in the transport process per se, but probably serve as receptors to trigger or initiate translocation of the solute through the membrane by binding to external sites of the integral membrane proteins of the efflux system.

In addition at least some solute-binding proteins function in the initiation of sensory transduction pathways.

On the basis of sequence similarities, the vast majority of these solute-binding proteins can be grouped [1] into eight families of clusters, which generally correlate with the nature of the solute bound.

Family 3 groups together specific amino acids and opine-binding periplasmic proteins and a periplasmic homolog with catalytic activity:

-Histidine-binding protein (gene hisJ) of Escherichia coli and related bacteria. An homologous lipoprotein exists in Neisseria gonorrhoeae.

812

-Lysine/arginine/ornithine-binding proteins (LAO) (gene argT) of *Escherichia coli* and related bacteria are involved in the same transport system than hisJ. Both solute-binding proteins interact with a common membrane-bound receptor hisP of the binding protein dependent transport system HisQMP.

5 -Glutamine-binding proteins (gene glnH) of *Escherichia coli* and *Bacillus stearothermophilus*.

-Glutamate-binding protein (gene gluB) of *Corynebacterium glutamicum*.

-Arginine-binding proteins artI and artJ of *Escherichia coli*.

-Nopaline-binding protein (gene nocT) from *Agrobacterium tumefaciens*.

10 -Octopine-binding protein (gene occT) from *Agrobacterium tumefaciens*.

-Major cell-binding factor (CBF1) (gene: peb1A) from *Campylobacter jejuni*.

-*Bacteroides nodosus* protein aabA.

-Cyclohexadienyl/arogenate dehydratase of *Pseudomonas aeruginosa*, a periplasmic enzyme which forms an alternative pathway for phenylalanine biosynthesis.

15 -*Escherichia coli* protein fliY.

-*Vibrio harveyi* protein pathH.

-*Escherichia coli* hypothetical protein ydhW.

-*Bacillus subtilis* hypothetical protein yckB.

-*Bacillus subtilis* hypothetical protein yckK.

20 The signature pattern is located near the N-terminus of the mature proteins.

Consensus pattern G-[FYIL]-[DE]-[LIVMT]-[DE]-[LIVMF]-x(3)-[LIVMA]-[VAGC]-x(2)-[LIVMAGN]

Sequences known to belong to this class detected by the pattern ALL.

25

[1] Tam R., Saier M.H. Jr. Microbiol. Rev. 57:320-346(1993).

1012. Sec7 - Sec7 domain

The Sec7 domain is a guanine-nucleotide-exchange-factor (GEF) for the arf family [2].

30

Number of members: 32.

813

[1]Medline: 98169075. Structure of the Sec7 domain of the Arf exchange factor. ARNO. Cherfils J, Menetrey J, Mathieu M, Le Bras G, Robineau S, Beraud-Dufour S, Antonny B, Chardin P; Nature 1998;392:101-105.

[2]Medline: 97100951. A human exchange factor for ARF contains Sec7- and pleckstrin-homology domains. Chardin P, Paris S, Antonny B, Robineau S, Beraud-Dufour S, Jackson CL, Chabre M. Nature 1996;384:481-484.

1013. SecA_protein. SecA protein, amino terminal region

SecA protein binds to the plasma membrane where it interacts with proOmpA to support translocation of proOmpA through the membrane. SecA protein achieves this translocation, in association with SecY protein, in an ATP dependent manner. SecA possesses the ATPase activity. The carboxyl terminus has similarity with the helicase carboxyl terminus. See Ribosomal_L5. Number of members: 45.

[1]Medline: 98309858. Amino-terminal region of SecA is involved in the function of SecE for protein translocation into Escherichia coli membrane vesicles. Mori H, Sugiyama H, Yamanaka M, Sato K, Tagaya M, Mizushima S; J Biochem (Tokyo) 1998;124:122-129.

[2]Medline: 89251629. SecA protein hydrolyzes ATP and is an essential component of the protein translocation ATPase of Escherichia coli. Lill R, Cunningham K, Brundage LA, Ito K, Oliver D, Wickner W; EMBO J 1989;8:961-966.

1014. Seedstore_2S - 2S seed storage family

Members of this family are composed of two chains (both included in the alignment), these are co-translated and later cleaved. The two chains are disulphide linked together. Number of members: 27.

[1]Medline: 97121264. 1H NMR assignment and global fold of napin BnIb, a representative 2S albumin seed protein. Rico M, Bruix M, Gonzalez C, Monsalve RI, Rodriguez R; Biochemistry 1996;35:15672-15682.

1015. Smr - Smr domain

This family includes the Smr (Small MutS Related) proteins, and the C-terminal region of the MutS2 protein. It has been suggested that this domain interacts with the MutS1 Swiss:P23909

protein in the case of Smr proteins and with the N-terminal MutS related region of MutS2
Swiss:P94545 [1]. Number of members: 14.

[1]Medline: 10431172. Smr: a bacterial and eukaryotic homologue of the C-terminal region
of the MutS2 family. Moreira D, Philippe H; Trends Biochem Sci 1999;24:298-300.

1016. (SSF) Sodium:solute symporter family signatures and profile

PROSITE: PDOC00429. PROSITE cross-reference(s)PS00456; NA_SOLUT_SYMP_1
PS00457; NA_SOLUT_SYMP_2 PS50283; NA_SOLUTE_SYMP_3

It has been shown [1,2] that integral membrane proteins that mediate the intake of a
wide variety of molecules with the concomitant uptake of sodium ions (sodium symporters)
can be grouped, on the basis of sequence and functional similarities into a number of distinct
families. One of these families is known as the sodium:solute symporter family (SSF) and
currently consists of the following proteins:

- Mammalian Na⁺/glucose co-transporter.
- Mammalian Na⁺/myo-inositol co-transporter.
- Mammalian Na⁺/nucleoside co-transporter.
- Mammalian Na⁺/neutral amino acid co-transporter.
- Escherichia coli Na⁺/proline symporter (gene putP).
- Escherichia coli Na⁺/pantothenate symporter (gene panF).
- Escherichia coli hypothetical protein yidK.
- Escherichia coli hypothetical protein yjcG.
- Bacillus subtilis hypothetical protein ywcA (ipa-31R).

These integral membrane proteins are predicted to comprise at least ten membrane
spanning domains. Two conserved regions were selected as signature patterns; the first one is
located in the fourth transmembrane region and the second one in a loop between two
transmembrane regions in the C-terminal part of these proteins.

Consensus pattern[GS]-x(2)-[LIY]-x(3)-[LIVMFYWSTAG](10)-[LIY]-[TAV]-x(2)-G-G-
[LMF]-x-[SAP]. Sequences known to belong to this class detected by the patternALL.

Consensus pattern[GAST]-[LIVM]-x(3)-[KR]-x(4)-G-A-x(2)-[GAS]-[LIVMGS]-[LIVMW]-
[LIVMGAT]-G-x-[LIVMGA] Sequences known to belong to this class detected by the
patternALL, except for E.coli yidK.

Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.

- 5 [1]Reizer J., Reizer A., Saier M.H. Jr. Res. Microbiol. 141:1069-1072(1991).
[2]Reizer J., Reizer A., Saier M.H. Jr. Biochim. Biophys. Acta 1197:133-136(1994).

1017. SurE - Survival protein SurE

E. coli cells with the surE gene disrupted are found to survive poorly in stationary phase [1].

- 10 It is suggested that SurE may be involved in stress response. Yeast also contains a member of the family Swiss:P38254. Swiss:P30887 can complement a mutation in acid phosphatase, suggesting that members of this family could be phosphatases. Number of members: 17.

[1]Medline: 95014035. A new gene involved in stationary-phase survival located at 59 minutes on the Escherichia coli chromosome. Li C, Ichikawa JK, Ravetto JJ, Kuo HC, Fu JC, Clarke S; J Bacteriol 1994;176:6015-6022.

[2]Medline: 93046805. Complementation of Saccharomyces cerevisiae acid phosphatase mutation by a genomic sequence from the yeast Yarrowia lipolytica identifies a new phosphatase. Treton BY, Le Dall MT, Gaillardin CM; Curr Genet 1992;22:345-355.

1018. Synuclein - Synuclein

There are three types of synucleins in humans, these are called alpha, beta and gamma.

Alpha synuclein has been found mutated in families with autosomal dominant Parkinson's disease. A peptide of alpha synuclein has also been found in amyloid plaques in Alzheimer's patients. Number of members: 12.

[1]Medline: 98424410. The synuclein family. Lavedan C; Genome Res 1998;8:871-880.

1019. (T-box) T-box domain signatures

- 30 PROSITE: PDOC00972. PROSITE cross-reference(s) PS01283; TBOX_1 PS01264;
TBOX_2

A number of eukaryotic DNA-binding proteins contain a domain of about 170 to 190 amino acids known as the T-box domain [1,2,3] and which probably binds DNA. The T-box

has first been found in the mice T locus (Brachyury) protein, a transcription factor involved in mesoderm differentiation. It has since been found in the following proteins:

-Vertebrate and invertebrate homologs of the T protein.

-Mammalian proteins TBX1 to TBX6.

5 -Mammalian protein TBR1 which is expressed specifically in brain.

-Xenopus laevis eomesodermin (eomes).

-Xenopus laevis Vegt (or Antipodean), a transcription factor that activates the expression of wnt-8, eomes and Brachyury.

-Chicken TbxT.

10 -Drosophila protein optomotor-blind (omb).

-Drosophila protein brachyenteron (byn) (also known as Trg), which is required for the specification of the hindgut and anal pads.

-Drosophila protein H15.

-Caenorhabditis elegans protein tbx-12.

5 -Caenorhabditis elegans hypothetical proteins F21H11.3, F40H6.4, T07C4.2, T07C4.6 and ZK177.10.

Two conserved regions were selected as signature patterns for the T-domain. The first region corresponds to the N-terminal of the domain and the second one to the central part.

Consensus pattern L-W-x(2)-[FC]-x(3,4)-[NT]-E-M-[LIV](2)-T-x(2)-G-[RG]-[KRQ]

Sequences known to belong to this class detected by the pattern ALL, except for C.elegans ZK177.10.

Consensus pattern [LIVMYW]-H-[PADH]-[DEN]-[GS]-x(3)-G-x(2)-W-M-x(3)-[IVA]-x- F

Sequences known to belong to this class detected by the pattern ALL, except for C.elegans

25 tbx-12, ZK177.10 and Drosophila H15.

[1] Bollag R.J., Siegfried Z., Cebra-Thomas J.A., Garvey N., Davison E.M., Silver L.M. Nat. Genet. 7:383-389(1994).

[2] Agulnik S.I., Garvey N., Hancock S., Ruvinsky I., Chapman D.L., Agulnik I., Bollag R.J.,

30 Papaioannou V.E., Silver L.M. Genetics 144:249-254(1996).

[3] Papaioannou V.E. Trends Genet. 13:212-213(1997).

1020. Toprim - Toprim domain

817

This is a conserved region from DNA primase. This corresponds to the Toprim domain common to DnaG primases, topoisomerases, OLD family nucleases and RecR proteins [1]. Both DnaG motifs IV and V are present in the alignment, the Dx D (V) motif may be involved in Mg²⁺ binding and mutations to the conserved glutamate (IV) completely abolish DnaG type primase activity [1]. DNA primase EC:2.7.7.6 is a nucleotidyltransferase it synthesizes the oligoribonucleotide primers required for DNA replication on the lagging strand of the replication fork; it can also prime the leading strand and has been implicated in cell division [2]. Number of members: 133.

[1]Medline: 98391745. Toprim--a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. Aravind L, Leipe DD, Koonin EV; Nucleic Acids Res 1998;26:4205-4213.

[2]Medline: 97368180. Cloning and analysis of the dnaG gene encoding Pseudomonas putida DNA primase. Szafranski P, Smith CL, Cantor CR; Biochim Biophys Acta 1997;1352:243-248.

[3]Medline: 94124015. The Haemophilus influenzae dnaG sequence and conserved bacterial primase motifs. Versalovic J, Lupski JR; Gene 1993;136:281-286.

1021. TraB - TraB family

pAD1 is a hemolysin/bacteriocin plasmid originally identified in Enterococcus faecalis DS16. It encodes a mating response to a peptide sex pheromone, cAD1, secreted by recipient bacteria. Once the plasmid pAD1 is acquired, production of the pheromone ceases--a trait related in part to a determinant designated traB. However a related protein is found in C. elegans Swiss:Q94217, suggesting that members of the TraB family have some more general function. Number of members: 12.

[1]Medline: 94302142. Characterization of the determinant (traB) encoding sex pheromone shutdown by the hemolysin/bacteriocin plasmid pAD1 in Enterococcus faecalis. An FY, Clewell DB; Plasmid 1994;31:215-221.

1022. (Transpo_mutator) Transposases, Mutator family, signature
PROSITE: PDOC00770. PROSITE cross-reference(s) PS01007;
TRANSPOSASE_MUTATOR

Autonomous mobile genetic elements such as transposon or insertion sequences (IS) encode an enzyme, called transposase, required for excising and inserting the mobile element. On the basis of sequence similarities, transposases can be grouped into various families. One of these families has been shown [1,2,3,E1] to consist of transposases from the following elements:

- Mutator from Maize.
- Is1201 from *Lactobacillus helveticus*.
- Is905 from *Lactococcus lactis*.
- Is1081 from *Mycobacterium bovis*.
- Is6120 from *Mycobacterium smegmatis*.
- Is406 from *Pseudomonas cepacia*.
- IsRm3 from *Rhizobium meliloti*.
- IsRm5 from *Rhizobium meliloti*.
- Is256 from *Staphylococcus aureus*.
- IsT2 from *Thiobacillus ferrooxidans*.

The maize Mutator transposase (MudrA) is a protein of 823 amino acids; the bacterial transposases listed above are proteins of 300 to 420 amino acids. These proteins contain a conserved domain of about 130 residues; a signature pattern was derived from the most conserved part of this domain.

Consensus pattern D-x(3)-G-[LIVMF]-x(6)-[STAV]-[LIVMFYW]-[PT]-x-[STAV]-x(2)-[QR]-x-C-x(2)-H. Sequences known to belong to this class detected by the pattern ALL.

[1] Eisen J.A., Benito M.-I., Walbot V. *Nucleic Acids Res.* 22:2634-2636(1994).

[2] Guilhot C., Gicquel B., Davies J., Martin C. *Mol. Microbiol.* 6:107-113(1992).

[3] Wood M.S., Byrne A., Lessie T.G. *Gene* 105:101-105(1991).

1023. Transposase_8 - Transposase

Transposase proteins are necessary for efficient DNA transposition. This family consists of various *E. coli* insertion elements and other bacterial transposases some of which are members of the IS3 family. Number of members: 58.

[1]Medline: 97324595. Genetic organization and transposition properties of IS511. D. A. Mullin, D. L. Zies, A. H. Mullin, N. Caballera & B. Ely; Mol Gen Genet 1997;254:456-463.

[2]Medline: 97128810. The use of an improved transposon mutagenesis system for DNA sequencing leads to the characterization of a new insertion sequence of *Streptomyces lividans* 66. J. Fischer, H. Maier, P. Viell & J. Altenbuchner; Gene 1996;180:81-89.

[3]Medline: 97074647. Identification and nucleotide sequence of *Rhizobium meliloti* insertion sequence ISRM6, a small transposable element that belongs to the IS3 family. S. Zekri & N. Toro; Gene 1996;175:43-48.

1024. tRNA_int_endo - tRNA intron endonuclease

Members of this family cleave pre tRNA at the 5' and 3' splice sites to release the intron
EC:3.1.27.9. Number of members: 8.

[1]Medline: 97344075. Properties of *H. volcanii* tRNA intron endonuclease reveal a relationship between the archaeal and eucaryal tRNA intron processing systems. Kleman-Leyer K, Armbruster DW, Daniels CJ; Cell 1997;89:839-847.

1025. Urease - Urease signatures

PROSITE: PDOC00133PROSITE cross-reference(s) PS01120; UREASE_1 PS00145;
UREASE_2

Urease (EC 3.5.1.5) is a nickel-binding enzyme that catalyzes the hydrolysis of urea to carbon dioxide and ammonia [1]. Historically, it was the first enzyme to be crystallized (in 1926). It is mainly found in plant seeds, microorganisms and invertebrates. In plants, urease is a hexamer of identical chains. In bacteria [2], it consists of either two or three different subunits (alpha, beta and gamma).

Urease binds two nickel ions per subunit; four histidine, an aspartate and a carbamated-lysine serve as ligands to these metals; an additional histidine is involved in the catalytic mechanism [3].

As signatures for this enzyme, a region that contains two histidine that bind one of the nickel ions and the region of the active site histidine was selected.

Consensus pattern T-[AY]-[GA]-[GAT]-[LIVM]-D-x-H-[LIVM]-H-x(3)-P [The two H's bind nickel]. Sequences known to belong to this class detected by the patternALL.

Consensus pattern[LIVM](2)-[CT]-H-[HN]-L-x(3)-[LIVM]-x(2)-D-[LIVM]-x-F-A [H is the active site residue]. Sequences known to belong to this class detected by the patternALL.

[1]Takishima K., Suga T., Mamiya G. Eur. J. Biochem. 175:151-165(1988).

5 [2]Mobley H.L.T., Husinger R.P. Microbiol. Rev. 53:85-108(1989).

[3]Jabri E., Carr M.B., Hausinger R.P., Karplus P.A. Science 268:998-1004(1995).

1026. Urease_beta - Urease beta subunit.

This subunit is known as alpha in Helicobacter. Number of members: 35.

10

[1]Medline: 95273988. The crystal structure of urease from Klebsiella aerogenes. Jabri E, Carr MB, Hausinger RP, Karplus PA; Science 1995;268:998-1004.

1027. UvrD-helicase - UvrD/REP helicase

5 The Rep family helicases are composed of four structural domains. The Rep family function as dimers. REP helicases catalyse ATP dependent unwinding of double stranded DNA to single stranded DNA. Swiss:P23478, Swiss:P08394 have large insertions near to the carboxy-terminus relative to other members of the family. Number of members: 52.

20 [1] Medline: 97433075. Major domain swiveling revealed by the crystal structures of complexes of E. coli Rep helicase bound to single-stranded DNA and ADP. Korolev S, Hsieh J, Gauss GH, Lohman TM, Waksman G; Cell 1997;90:635-647.

1028. V-type ATPase 116kDa subunit family (V_ATPase_sub_a)

25

This family consists of the 116kDa V-type ATPase (vacuolar (H⁺)-ATPases) subunits, as well as V-type ATP synthase subunit i. The V-type ATPases family are proton pumps that acidify intracellular compartments in eukaryotic cells for example yeast central vacuoles, clathrin-coated and synaptic vesicles. They have important roles in membrane trafficking processes [1]. The 116kDa subunit (subunit a) in the V-type ATPase is part of the V₀ functional domain responsible for proton transport. The a subunit is a transmembrane glycoprotein with multiple putative transmembrane helices. It has a hydrophilic amino terminal and a hydrophobic carboxy terminal [1,2]. It has roles in proton transport and

30

821

assembly of the V-type ATPase complex [1,2]. This subunit is encoded by two homologous gene in yeast VPH1 and STV1 [2].

Number of members: 27

- 5 [1] Forgac M; Medline: 99240666 "Structure and properties of the vacuolar (H⁺)-ATPases." J Biol Chem 1999;274:12951-12954.
[2] Forgac M; Medline: 99270697 "Structure and properties of the clathrin-coated vesicle and yeast vacuolar V-ATPases." J Bioenerg Biomembr 1999;31:57-65.

10 1029. Viral (Superfamily 1) RNA helicase (Viral_helicase1)
Number of members: 260

[1] Koonin EV, Dolja VV; Medline: 94094568 "Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences." Crit Rev Biochem Mol Biol 1993;28:375-430.

1030. Vesicular monoamine transporter (VMAT)

This family consists of various vesicular amine transporters with 12 transmembrane helices. These included vesicular acetylcholine transporters (VACHT) [3], and vesicular monoamine transporters (VMATs) [1,2] isoforms 1 adrenal and 2 brain (VMAT1 and VMAT2).

These proteins transport biogenic amines into synaptic vesicles or chromaffin granules [4]. VMATs pack monoamine neurotransmitters into secretory vesicles for regulated exocytotic release, they also protect against the parkinsonian neurotoxins MPP⁺ by transporting it into vesicles preventing it from acting on mitochondria [1].

Also in the family is C. elegans UNC-17 a putative vesicular acetylcholine transporter mutations in UNC-17 cause impaired neuromuscular function, giving rise to jerky or uncoordinated movement, [4].

Number of members: 15

[1] Krantz DE, Peter D, Liu Y, Edwards RH; Medline: 97197857 "Phosphorylation of a vesicular monoamine transporter by casein kinase II." J Biol Chem 1997;272:6752-6759.

[2] Erickson JD, Varoqui H, Schafer MK, Modi W, Diebler MF, Weihe E, Rand J, Eiden LE, Bonner TI, Usdin TB; Medline: 94350930 "Functional identification of a vesicular acetylcholine transporter and its expression from a 'cholinergic' gene locus." J Biol Chem 1994;269:21929-21932.

[3] Erickson JD, Schafer MK, Bonner TI, Eiden LE, Weihe E; Medline: 96209876 "Distinct pharmacological properties and distribution in neurons and endocrine cells of two isoforms of the human vesicular monoamine transporter." Proc Natl Acad Sci U S A 1996;93:5166-5171.

[4] Alfonso A, Grundahl K, Duerr JS, Han HP, Rand JB; Medline: 3342494 "The Caenorhabditis elegans unc-17 gene: a putative vesicular acetylcholine transporter." Science 1993;261:617-619.

1031. WW/rsp5/WWP domain signature and profile. Cross-reference(s): PS01159; WW_DOMAIN_1; PS50020; WW_DOMAIN_2

The WW domain [1-4,E1] (also known as rsp5 or WWP) has been originally discovered as a short conserved region in a number of unrelated proteins, among them dystrophin, the gene responsible for Duchenne muscular dystrophy. The domain, which spans about 35 residues, is repeated up to 4 times in some proteins. It has been shown [5] to bind proteins with particular proline-motifs, [AP]-P-P-[AP]-Y, and thus resembles somewhat SH3 domains. It appears to contain beta-strands grouped around four conserved aromatic positions; generally Trp. The name WW or WWP derives from the presence of these Trp as well as that of a conserved Pro. It is frequently associated with other domains typical for proteins in signal transduction processes.

Proteins containing the WW domain are listed below.

--Dystrophin, a multidomain cytoskeletal protein. Its longest alternatively spliced form consists of an N-terminal actin-binding domain, followed by 24 spectrin-like repeats, a cysteine-rich calcium-binding domain and a C-terminal globular domain. Dystrophin form tetramers and is thought to have multiple functions including involvement in membrane stability, transduction of contractile forces to the extracellular environment and organization

of membrane specialization. Mutations in the dystrophin gene lead to muscular dystrophy of Duchenne or Becker type. Dystrophin contains one WW domain C-terminal of the spectrin-repeats.

--Utrophin, a dystrophin-like protein of unknown function.

5 --Vertebrate YAP protein is a substrate of an unknown serine kinase. It binds to the SH3 domain of the Yes oncoprotein via a proline-rich region. This protein appears in alternatively spliced isoforms, containing either one or two WW domains [6].

10 --Mouse NEDD-4 plays a role in the embryonic development and differentiation of the central nervous system. It contains 3 WW modules followed by a HECT domain. The human ortholog contains 4 WW domains, but the third WW domain is probably spliced resulting in an alternate NEDD-4 protein with only 3 WW modules [3].

--Yeast RSP5 is similar to NEDD-4 in its molecular organization. It contains an N-terminal C2 domain (see <PDOC00380>), followed by a histidine-rich region, 3 WW domains and a HECT domain.

15 --Rat FE65, a transcription-factor activator expressed preferentially in liver. The activator domain is located within the N-terminal 232 residues of FE65, which also contain the WW domain.

20 --Yeast ESS1/PTF1, a putative peptidyl prolyl cis-trans isomerase from family ppiC (see <PDOC00840>). A related protein, dodo (gene dod) exists in Drosophila and in mammals (gene PIN1).

--Tobacco DB10 protein. The WW domain is located N-terminal to the region with similarity to ATP-dependent RNA helicases.

--IQGAP, a human GTPase activating protein acting on ras. It contains an N-terminal domain similar to fly muscle mp20 protein and a C-terminal ras GTPase activator domain.

25 --Yeast pre-mRNA processing protein PRP40, Caenorhabditis elegans ZK1098.1 and fission yeast SpAC13C5.02 are related proteins with similarity to MYO2-type myosin, each containing two WW-domains at the N-terminus.

--Caenorhabditis elegans hypothetical protein C38D4.5, which contains one WW module, a PH domain (see <PDOC50003>) and a C-terminal phosphatidylinositol 3-kinase domain.

30 --Yeast hypothetical protein YFL010c.

For the sensitive detection of WW domains, a profile was developed which spans the whole homology region as well as a pattern.

Description of pattern(s) and/or profile(s):

Consensus pattern W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTOCR]-[FYW]-x(2)-P.

5

[1] Bork P., Sudol M. Trends Biochem. Sci. 19:531-533(1994).

[2] Andre B., Springael J.Y. Biochem. Biophys. Res. Commun. 205:1201-1205(1994).

[3] Hofmann K.O., Bucher P. FEBS Lett. 358:153-157(1995).

[4] Sudol M., Chen H.I., Bougeret C., Einbond A., Bork P. FEBS Lett. 369:67-71(1995).

10 [5] Chen H.I., Sudol M. Proc. Natl. Acad. Sci. U.S.A. 92:7819-7823(1995).

[6] Sudol M., Bork P., Einbond A., Kastury K., Druck T., Negrini M., Huebner K., Lehman D. J. Biol. Chem. 270:14733-14741(1995).

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

1032. XPA protein signatures, cross-reference(s): XPA_1 PROSITE PS00752;

PS00753;XPA_2.

Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair. There are a minimum of seven genetic complementation groups involved in this pathway: XP-A to XP-G. XP-A is the most severe form of the disease and is due to defects in a 30 Kd nuclear protein called XPA (or XPAC) [2].

The sequence of the XPA protein is conserved from higher eukaryotes [3] to yeast (gene RAD14) [4]. XPA is a hydrophilic protein of 247 to 296 amino-acid residues which has a C4-type zinc finger motif in its central section.

Two signature were developed patterns for XPA proteins. The first corresponds to the zinc finger region, the second to a highly conserved region located some 12 residues after the zinc finger region.

Consensus pattern C-x-[DE]-C-x(3)-[LIVMF]-x(1,2)-D-x(2)-L-x(3)-F-x(4)-C-x(2)-C

Consensus pattern [LIVM](2)-T-[KR]-T-E-x-K-x-[DE]-Y-[LIVMF](2)-x-D-x-[DE]

[1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).

[2] Miura N., Miyamoto I., Asahina H., Satokata I., Tanaka K., Okada Y. J. Biol. Chem. 266:19786-19789(1991).

5 [3] Shimamoto T., Kohno K., Tanaka K., Okada Y. Biochem. Biophys. Res. Commun. 181:1231-1237(1991).

[4] Bankmann M., Prakash L., Prakash S. Nature 355:555-558(1992).

1033. YCF9

10 This family consists of the hypothetical protein product of the YCF9 gene from chloroplasts and cyanobacteria. Number of members: 16

1034. (DUF15)

15 It is highly conserved between eubacteria and eukaryotes.

Number of members: 30

1035. Luminal portion of Cytochrome b559, alpha (gene psbE) subunit. (cytochr_b559a)

20 This family is the luminal portion of cytochrome b559 alpha chain, matches to this family should be accompanied by a match to the cytochr_b559 family also. The Prosite pattern matches the transmembrane region of the cytochrome b559 alpha and beta subunits.

Number of members: 16

25

A. Asparaginase 2

30 Asparaginase II (L-asparagine aminohydrolase II) is an extracellular protein that may be associated with the cell wall and whose expression is affected by the availability of nitrogen. Asparaginase II catalyzes the reaction of L-Asparagine + H₂O = L-Aspartate + NH₃. As many leukemias have high requirements for aspartic acid, asparaginase II proteins are useful

as reagents for screening compounds for activity as leukemia chemotherapy products. Asparaginase II protein can also be over- or under-expressed to alter amino acid content in plant tissues or to modify nitrogen fixation and/or nitrogen metabolism in plants.

- 5 Ref: Bon et al. (1997) Appl Biochem Biotechnol 63-65: 203-12

B. Chloroa b-bind

Chlorophyll a-b binding proteins are located in the thylakoid membranes of the chloroplast and bind chlorophyll a and chlorophyll b, thereby triggering a chemical reaction (photosynthesis). These proteins are useful in controlling the rate, efficiency and/or output of photosynthesis. Overexpression of chlorophyll a-b binding proteins is expected to increase the rate of photosynthesis.

- 10 Ref: Leutwiler et al. (1986) Nucleic Acids Res 14: 4051-64
Brandt et al. (1992) Plant Mol Biol 19: 699-703

C. DMRL synthase

DMRL Synthase (6,7-Dimethyl-8-Ribityllumazine Synthase) catalyzes the last step in riboflavin (Vitamin B₂) synthesis, condensing 5-amino-6-(1'-D)-ribityl-amino-2,4(1H, 3H)-Pyrimidinedione with L-3,4-Dihydroxy-2-Butanone 4-Phosphate producing 6,7-Dimethyl-8-(1-D-Ribityl)Luminazine. The enzyme forms a homopentamer. Engineering of these proteins or those with homologous sequences/structures may allow control of the amounts of vitamin B₂ available in plants and/or accumulation of pigment, as well as altering reactions requiring hydrogen ion carriers/transmitters.

- 25 Ref: Garcia-Ramirez et al. (1995) J Biol Chem **270**: 23801-7

D. E1_N

These proteins are ATP-dependent DNA helicases that are required for initiation of viral DNA replication. They form a complex with the viral E2 protein. The E1-E2 complex binds

to the replication origin that contains binding sites for both proteins. The majority of sequences known for this group of proteins are from various papillomaviruses, a type of double stranded DNA virus. In plants, the prototype double stranded DNA virus is Cauliflower Mosaic virus (CaMV). Manipulation of these proteins, especially to produce variant proteins that form non-productive complexes, enables production of plants that are resistant to infection by double stranded DNA viruses.

Ref: Yang et al. (1993) PNAS USA **90**: 5086-90

Ustav and Stenlund (1991) EMBO J **10**: 449-57

Callaway et al. (1996) Mol Plant Microbe Interact **9**: 810-8

E. EF1 G

Elongation Factor-1 is composed of four subunits: alpha, beta, delta and gamma. Gamma subunits are presumed to play a role in anchoring the complex to other cellular components. Studies of EF-1 genes in plants suggests that different forms of the EF-1 subunits may be expressed in particular organs or in response to stress. Manipulation of the activity of these proteins, either by altered expression level or by structural mutation, may result in the accumulation of a particular protein in a chosen organ or allow production of particular proteins during stress conditions.

Ref: Kinzy et al. (1994) NAR **22**: 2703-7

Dunn et al. (1993) Plant Mol Biol **23**: 221-5

Aguilar et al. (1991) Plant Mol Biol **17**: 351-60

F. ENV polyprotein

This family comprises the envelope or coat proteins known from a number of different retroviruses. In mammalian species, retroviruses are responsible for diseases such as leukemia and HIV. In plants, retroviruses are known in both monocot (e.g. Zeon-1) and dicot (e.g. Arabidopsis and tobacco) species and have been shown to induce mutant alleles at new loci. Engineering of plant ENV proteins may allow mobilization or targeting of endogenous

or introduced retroviruses, in essence generating a new method for mutant production, gene tagging and the like.

Ref: Mamoun et al (1990) J Virol 64: 4180-8
5 Grandbastien et al. (1989) Nature 337: 376-80
Wright and Voytas (1998) Genetics 149: 703-15

G. Glycosyl_hydr9

10 Proteins having this domain (previously known as the glycosyl hydrolase family 5 domain) catalyze the endohydrolysis of 1,4- β -D-glucosidic linkages in cellulose. Numerous plant proteins with this domain exist and are expressed in an organ specific manner. They are involved in the fruit ripening process, in cell elongation and plant reproduction. Modulation
15 of the activity of these proteins, either by over- or under-expression or by mutation of the polypeptide, could be used to affect post-harvest physiology (e.g. rate of ripening) or for engineering reproductive sterility.

Ref: Giorda et al. (1990) Biochemistry 29: 7264-9
20 Tucker et al. (1988) Plant Physiol 88: 1257-62
Shani et al. (1997) 43: 837-42
Milligan and Gasser (1995) Plant Mol Biol 28: 691-711

H. Glycosyl_hydr14

25 The β -amylases (family 14 of glycosyl hydrolases) catalyze the hydrolysis of 1,4- α -glucosidic linkages in polysaccharides and remove successive maltose units from the non-reducing ends of the chains. Mutants of β -amylase in Arabidopsis exhibited altered degradation of starch throughout the diurnal cycle. In addition, the mutant phenotypes
30 indicated that these enzymes not only affect carbohydrate metabolism/catabolism, but also influence the amount of pigment stored within particular cells. Manipulation of the β -amylase genes enables control of plant pigmentation (for example, fibre pigment in cotton) as well as carbohydrate synthesis and degradation.

Ref: Zeeman et al. (1998) Plant J 15: 357-65
Hirano and Nakamura (1997) Plant Physiol 114: 5675-82
Kitamoto et al. (1988) J Bacteriol 170: 5848-54

5

I. Glycosyl_hydr15

Glycosyl hydrolases from family 15 (such as 1,4-Alpha-D-Glucan glucohydrolase,) catalyze the hydrolysis of terminal 1,4-linked alpha-D-glucose residues successively from the non-reducing ends of the chains resulting in the release of β -D-Glucose. In plants these proteins have been tied to the mobilization of the xyloglucan stored in the cotyledonary cell walls. Proteins such as these could be varied to affect the rate of plant growth (for example during germination), storage and/or use of glucose and other sugars by plant tissues and alteration of the properties, such as elasticity, of plant cell walls.

Ref: Crombie et al. (1998) Plant J 15: 27-38
Hata et al. (1991) Agric Biol Chem 55: 941-9

J. Glycosyl_hydr20

Members of the family 20 glycosyl hydrolases catalyze the hydrolysis of terminal non-reducing N-acetyl-D-hexosamine residues in N-acetyl- β -D-hexosaminides. N-acetyl- β -glucosaminidase belongs to this family and exists in several different forms (consisting of various combinations of alpha and beta chains) depending on the organism. Family 20 glycosyl hydrolases have been implicated in lysosomal storage diseases (such as Sandhoff disease) and glycogen storage disease in humans. These types of proteins are also responsible for the hydrolysis of chitin. In plants, these proteins could be useful in controlling carbohydrate catabolism, thereby influencing the amount of sugars available for storage and/or use in other metabolic pathways. In addition, it is possible that such proteins could be used to engineer an endogenous insect protection mechanism, e.g. by secretion of a chitin-hydrolyzing composition by the plant.

Ref: Graham et al (1988) J Biol Chem 263: 16823-9

O'Dowd et al. (1988) Biochemistry 27: 5216-26

K. HMG box

5

The HMG box is a novel type of DNA-binding domain found in a diverse group of proteins. Numerous plant proteins contain this domain, such as the HMG1/2-like proteins. The expression of some of these HMG proteins appears to be regulated by circadian rhythms and in a light dependent manner, occurring at higher levels in roots, for example and lower levels in light-grown tissues such as cotyledons. Generally, HMG proteins are thought to influence transcription regulation. In plants, HMGs are believed to have a role in maintaining patterns of circadian-regulated expression for other genes, suggesting that these proteins could be exploited to control growth and development.

10

Ref: Laudet et al. (1993) Nucleic Acids Res 21: 2493-501

Zheng et al. (1993) Plant Mol Biol 23: 813-23

Grasser et al. (1993) Plant Mol Biol 23: 619-25

L. IL2

20

Interleukin-2 (IL-2) is produced in mammals by T cells in response to antigenic or mitogenic stimulation and is crucial for proper regulation and functioning of the immune response. IL-2 is capable of stimulating B cells, monocytes, lymphokine-activated killer cells, natural killer cells and glioma cells. Plant extracts have also been shown to stimulate the immune system (for example, mistletoe therapy for human cancer). It is known that IL-2 is involved in feedback inhibition pathways that impact the inflammatory response as well as the growth inhibition of tumor reactive T cells. Plant proteins containing IL-2-like sequences are useful as immunity-based therapeutics, acting in a manner similar to IL-2 in mammals.

25

Ref: Heike et al. (1997) Scand J Immunol 45: 221-6

Ariel et al. (1998) J Immunol 161: 2465-72

Schink (1997) Anticancer Drugs 8 Suppl 1: S47-51

30

M. Oxidored_FMN

NADPH dehydrogenases catalyze the reaction $\text{NADPH} + \text{acceptor} = \text{NADP}(+) + \text{reduced acceptor}$. One member of this family is yeast "old yellow enzyme" (OYE) and is thought to be involved in oxylipin metabolism. A second yeast family member is a protein that binds estrogen binding protein (EBP) in addition to exhibiting oxidoreductase activity. An Arabidopsis homolog to OYE has been described and estrogen binding proteins in plants have been reported. Plant proteins from this class have the potential to be used to modify lipid metabolism/catabolism. These proteins may also have use as therapeutics for breast and prostate cancer, and other abnormal growth in steroid-sensitive tissues.

Ref: Baker et al. (1998) Proc Soc Exp Biol Med 217: 317-21
Schaller and Weiler (1997) J Biol Chem 272: 28066-72
Mandani et al. (1994) PNAS USA 91: 922-6

N. Oxidored_q2

The NADH-plastoquinone oxidoreductases catalyze the reaction $\text{NADH} + \text{plastoquinone} = \text{NAD}(+) + \text{plastoquinol}$. In plants these reactions occur in the chloroplast and are believed to participate in a chloroplast respiratory system. Here, the NDH complex is postulated to act as a valve to remove excess reduction equivalents in the chloroplasts. Manipulation of these proteins may improve the rate or efficiency of photosynthesis.

Ref: Burrows et al. (1998) EMBO J 17: 868-76
Kofer et al (1998) Mol Gen Genet 258: 166-73
Maier et al. (1995) J Mol Biol 251: 614-28

O. PABP

Polyadenylate binding proteins bind the poly (A) tail of mRNA. Plants, as exemplified by Arabidopsis, contain numerous PABP genes that are expressed in an organ-specific manner. For example, PABP2 is functional in roots and shoots, while PABP5 is expressed predominantly in immature flowers. The PABP proteins are implicated in numerous aspects

of posttranscriptional regulation including mRNA turnover and translational initiation. Control of activity of PABP proteins provides the ability to control the expression of various genes in particular organs during development.

- 5 Ref: Hilson et al (1993) Plant Physiol 103: 525-33
Belostotsky and Meagher (1993) PNAS USA 90: 6686-90

P. Parvo coat

10 Parvoviruses are linear single-stranded DNA viruses that are encapsulated by three capsid proteins. Plants are susceptible to infection by single stranded DNA viruses such as Maize streak virus (MSV) and various Gemini viruses. The coat proteins in these plant viruses are critical to the virus life cycle within the plant. For example, the coat protein of MSV is thought to be involved in intra- and inter-cellular movement within the plant. Engineering of
5 proteins having similarity to parvoviral coat proteins, especially to produce proteins that interfere with maturation of the virus particle, enables the production of plants having better resistance to natural plant single-stranded DNA viruses.

- Ref: Liu et al. (1997) J Gen Virol 78: 1265-70
20 Rohde et al. (1990) Virology 176: 648-51

Q. Pkinase_C

Plant serine/threonine protein kinases possessing this domain are expressed in all tissues and
25 are known to undergo serine-specific autophosphorylation and specifically phosphorylate two ribosomal proteins, P14 and P16. During development, these proteins predominate during high metabolic activity in growing buds, root tips, leaf margins and germinating seeds. They are thought to be involved in the control of plant growth and development. In addition, two genes encoding proteins from this family have been described that help plant cells adapt
30 during cold or high salt stresses. Consequently, engineering Pkinase C proteins provides a way to control general growth/development of the plant as well as a means to provide endogenous protection against environmental stresses.

Ref: Zhang et al. (1994) J Biol Chem 269: 17586-92

Mizoguchi et al. (1995) FEBS Lett 358: 199-204

R. REV

5

The REV proteins act post-transcriptionally to relieve negative repression of GAG and ENV production in retroviruses such as Human Immunodeficiency Virus type I (HIV-1). Plants contain retrovirus-like viruses such as pararetroviruses and retrotransposons (i.e. transposons having long terminal repeats). Plant retrotransposons in particular have been used to create mutations at various loci, thereby permitting gene isolation, gene tagging and the like. Manipulation of plant REV proteins enables control of transposition frequencies of corresponding transposable elements and provides a new tool for genetic engineering of plants.

10

Ref: Sodroski et al. (1986) Nature 321: 412-7

Franchini et al. (1989) PNAS USA 86: 2433-7

Marquet et al. (1995) 77: 113-24

Grandbastien et al. (1989) Nature 337: 376-80

Wright and Voytas (1998) Genetics 149: 703-15

15

20

S. RuBisCo small

Ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCo) catalyzes the initial step in the C3 photosynthetic carbon reduction cycle, adding carbon dioxide to D-ribulose 1,5-bisphosphate to form two molecules of 3-phospho-D-glycerate. RuBisCo is comprised of two subunits, one large which is synthesized in the chloroplast, and one small which is synthesized in the cytoplasm and then transported in to the chloroplast. The expression of the small subunit of RuBisCo is light regulated. Manipulation of these proteins could increase the efficiency of photosynthesis or allow alterations in developmental timing.

25

30

Ref: Giuliano et al. (1988) PNAS USA 85: 7089-93

Dedonder et al. (1993) Plant Physiol 101: 801-8

T. Sialyltransf

Members of the CMP-N-acetylneuraminate- β -galactosamide- α -2,3-sialyltransferase family catalyze the following reaction:

5 CMP-N-acetylneuraminate + β -D-galactosyl-1,3-N-acetyl- α -D-galactosaminyl-R = CMP + α -N-acetylneraminyl-2,3- β -D-galactosyl-1,3-N-acetyl- α -D-galactosaminyl-R. These proteins are thought to be responsible for the synthesis of the sequence neurac- α -2,3-gal- β -1,3-galnac- found on sugar chains)-linked to threonine or serine and also as a terminal sequence on certain gangliosides in mammalian cells. In plants, glycosyltransferases in the
10 Golgi apparatus synthesize cell wall polysaccharides and elaborate the complex glycans of glycoproteins. Engineering of plant sialyltransferases allows targeting of proteins to particular cellular locations or enables the making of changes in cell wall structure.

Ref: Wee et al. (1998) Plant Cell 10: 1759-68

45 Lee et al. (1994) J Biol Chem 269: 10028-33

Kitagawa and Paulson (1994) J Biol Chem 269: 1394-401

U. Signal

20 Many plant proteins in this family contain sequences similar to those found in both components of the prokaryotic family of signal transducers known as the two-component systems. This suggests that activation may require a transfer of a phosphate group between the transmitter domain and the receiver domain. One family member in Arabidopsis appears to be involved in ethylene (a plant hormone) signal transduction. Other proteins in this family
25 appear to be involved in the regulation of gene transcription under conditions of environmental stress. Signal proteins can be exploited to affect plant growth and development and/or control plant responses to stress conditions such as cold, nutrient availability, etc.

Ref: Chang et al. (1993) Science 262: 539-44

30 Nagaya et al. (1993) Gene 131: 119-124

Gottfert et al. (1990) PNAS USA 87: 2680-4

V. vMSA

vMSA proteins are major surface antigens presenting on the envelope of various retroviruses. Surface antigens of retroviruses are often involved in tropism of the virus.

5 Plants contain retrovirus-like viruses such as pararetroviruses and retrotransposons (i.e. transposons having long terminal repeats). Plant retrotransposons in particular have been used to create mutants at various loci, thereby permitting gene isolation, gene tagging and the like. Manipulation of plant vMSA proteins enables control of tropism of plant retroviruses that might be used for genetic engineering tools, thus enabling targeting of the virus to
10 particular species and/or tissues of plants.

Ref: Okamoto et al. (1988) J Gen Virol 69: 2575-83

Grandbastien et al. (1989) Nature 337: 376-80

Wright and Voytas (1998) Genetics 149: 703-15

W. zf-CCCH

This family of proteins is defined by having two CX(8)CX(5)CX(3)H-type zinc finger domains. These proteins cover a broad range of functions. For example, the COP1 protein acts as a repressor of photomorphogenesis in darkness; light stimuli abolish this suppressive action. In addition, COP1 protein can function as a negative transcriptional regulator capable of direct interaction with components of the G-protein signaling pathway. As a second example, a zf-CCCH protein identified in Arabidopsis appears to be involved in the resistance to DNA damage induced by UV light and chemical DNA-damaging agents.

25 Overexpression of this class of proteins permits production of plants that are better suited to adverse environments. Manipulation of expression of zf-CCCH proteins functioning as transcriptional regulators, such as COP1, enables manipulation of some signal transduction pathways.

30 Ref: Pang et al. (1993) Nucleic Acids Res 21: 1647-53

Deng et al. (1992) Cell 71: 791-801

X. zf-RanBP

Proteins falling within this category contain many X-X-F-G and X-F-X-F-G repeats, and may contain RANBP1-like or PPIase domains. Plant proteins having domains similar to these include PAS1 and GMSTI. PAS1 has been shown to have dramatic developmental affects that appear to be correlated with both cell division and cell wall elongation. GMSTI has high identity to the yeast STI stress-inducible gene and has been shown to be heat inducible. Proteins such as these may be useful for controlling growth and form of development.

Ref: Vittorioso et al. (1998) Mol Cell Biol 18: 3034-43
Hernandez Torres et al. (1995) 27: 1221-6

Y. Peptidase M48.

Proteins belonging to this peptidase family are metalloproteases that bind zinc as a cofactor and are located in the membranes of the endoplasmic reticulum. They function in NH₂-terminal proteolytic processing, as shown for the yeast STE24 gene product. This gene is required for the correct processing of α -factor, a yeast pheromone. Family M48 peptidases also appear to be required for some prenylation reactions, mediating COOH-terminal CAAX processing. Prenylation reactions are believed to be involved in the regulation of protein-protein and protein-membrane interactions. As an example, RAS GTPase activity is regulated in part by localization to the inner side of the plasma membrane upon prenylation. In plants, proteins from this family could be involved in pollen-stigma interactions such as those mediating self-pollination vs. outcrossing, or could be members of several secondary metabolism pathways.

Ref: Fujimura-Kamada et al. (1997) J Cell Biol. 136: 271-85. Tam et al. (1998) J Cell Biol. 142: 635-49.

Z. DNA Pol Viral N

The DNA pol Viral N domain is located at the N-terminal region of DNA polymerase isolated from several retroid viruses such as the Cauliflower Mosaic Virus. The domain motif has also been found in numerous other species from humans to cyanobacteria. In these

organisms, this motif seems to be associated with two types of sequences; retrotransposons and mitochondrial genes. In the mitochondrial sequences this domain is potentially involved in the self-splicing conducted by group II introns. Various manipulations of this gene in plants allows control of the numerous retrotransposons endogenous to plant genomes or
 5 allows engineering of mitochondrial function, especially to increase efficiency of energy utilization by cells.

REF: Chapdelaine and Bonen (1991) Cell 65: 465-72
 Ferat and Miche (1993) Nature 364: 358-61
 10 Wilson et al. (1994) 368: 32-8
 Cambareri et al. (1994) 242: 658-65
 Gaardner et al. (1981) NAR 9: 2871-2888
 Cummings et al. (1990) Curr Genet 17: 375-402
 Hattori et al. (1986) Nature 321: 625-8

Aa. Calpain_inhib

This domain is found in calpastatin, an inhibitor protein specific for calpain. Calpain is a non-lysosomal calcium-dependent intracellular protease that appears to be involved in the dynamic changes of the cytoskeleton, especially actin-related structures, during early
 20 *Drosophila* embryogenesis [1]. Calpastatins co-exist in cells with calpains and the subcellular distribution of calpastatin is thought to be important to calpain regulation [2]. In plants calpains and calpastatins could be involved in embryogenesis and non-embryogenic organ reiteration. Mutations occurring in calpain inhibitor repeat domains would produce developmental abnormalities such as abnormal leaf, root or flower development.

Refs

- 1 Emori Y and Saigo K (1994) J Biol Chem 269: 25137-42.
- 2 Mellgren RL, Lane RD, Mericle MT (1989) Biochim Biophys Acta 999: 71-77.

Ab. chorismate_bind

Chorismate binding domains are present in plant anthranilate synthase (AS) genes. AS genes catalyze the first step in the biosynthesis of tryptophan by converting chorismate and L-glutamine to anthranilate, pyruvate and L-glutamate. Some of these genes are involved in

feedback inhibition by tryptophan [1] while some are feedback insensitive [2]. In Arabidopsis, two AS genes have overlapping, but different distributions. One of these AS genes is induced by wounding and bacterial pathogen infiltration [1]. Mutations in the chorismate binding domain would affect the production of tryptophan and could influence the plant's defense system. AS gene products can be used for *in vitro* synthesis of tryptophan and tryptophan derivatives.

Refs

- 1 Niyogi KK, Fink GR (1992) Plant Cell 4: 721-33.
- 2 Song HS, Brotherton JE, Gonzales RA, Wilholm JM (1998) Plant Physiol 117:533-43.

Ac. late protein L2

Papillomaviruses are encapsulated double stranded DNA viruses. Plants are susceptible to infection by double stranded DNA viruses such as Cauliflower Mosaic virus (CaMV). The coat proteins in these plant viruses are critical to the virus life cycle within the plant. For example, the coat protein of CaMV is thought to be involved in intra- and inter-cellular movement within the plant [1]. Engineering of proteins having similarity to papillomavirus coat proteins may enable the production of plants having better resistance to natural plant double stranded DNA viruses.

Refs

- 1 Thompson SR, Melcher U (1993) J Gen Virol 74: 1141-8.

Ad. Peptidase M41

Proteins belonging to this peptidase family are metalloproteases that bind zinc as a cofactor and are integral membrane proteins. They seem to be involved in the degradation of carboxy-terminal-tagged cytoplasmic proteins. In plants, these proteins are located in the thylakoid membranes of the chloroplasts, their expression is light regulated and they are thought to be involved in degradation of soluble stromal proteins and turn-over of thylakoid proteins [1]. Manipulation of expression and structure of these proteins would have effects on the efficiency of photosynthesis and the development of chloroplasts.

Refs

1 Lindahl M, Tabak s, Cseke L, Pichersky E, Andersson B, Adam Z (1996) J Biol Chem 271: 29329-34.

5 Ae. UPF0051

There is some evidence that, in plants, proteins in this family are involved in ATP synthesis in chloroplasts [1, 2]. Mutations in these proteins or altering their expression would affect the efficiency of photosynthesis and energy production.

10 Refs

1 Kostrzewa M, Zetsche K (1992) J Mol Biol 227: 961-70.

2 Kostrzewa M, Zetsche K (1993) Plant Mol Biol 23: 67-76

Af. E7

15 Papillomaviruses are encapsulated double stranded DNA viruses. The Papillomavirus early protein 7 (E7) is known as a potent immortalizing and transforming agent. Transformation by E7 is thought to be mediated by the physical association of E7 with cellular proteins regulating entry into the cell cycle [1]. The result is entry into the cell cycle and suppression of terminal differentiation in mammalian cells. Thus, engineering of proteins having
20 similarity to papillomavirus E7 protein enables the production of plants having altered cellular proliferation characteristics and possibly altered morphology. For example, overexpression of E7-like proteins would be expected to result in proliferation of cells of the tissue in which the E7 protein is expressed, perhaps with suppression of differentiation events. Thus, for example, overexpression of E7-like proteins in meristem cells can result in
25 taller plants and suppression of leafing and/or flowering.

Refs

1 Zwerschke W, Jansen-Durr P Adv Cancer Res 2000;78:1-29

30 Ag. Peptidase U7

This protein is known to be an integral membrane protein in the cyanobacterium Synechocystis where it functions to digest cleaved signal peptides [1]. This activity is necessary to maintain proper secretion of mature proteins across the membrane. In higher

plants this protein may be present in the plastid or chloroplast membranes where it would function by enabling protein movement into and out of the chloroplasts. Mutations in this protein would be expected to affect the development of plastids, including chloroplasts, or alter the energy transfer system within the chloroplasts, thereby affecting growth and development.

Refs

- 1 Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirose M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S (1996) DNA Res 3:109-36.

Ah. 5'-3' Exonuclease

The 5'-3' exonuclease domain is one found in bacterial DNA polymerases I and in yeast DNA repair enzymes such as Exonuclease I. Yeast Exo I is involved in mitotic recombination and also includes a domain that interacts with the mismatch repair protein MSH2. The 5'-3' exonuclease domain is also present in XPG DNA repair enzymes in humans and in yeast RAD9 protein. Defects in XPG proteins result in Xeroderma Pigmentosum. Thus defects in 5'-3' exonuclease domain-containing proteins in plants are expected to lead to defects in DNA repair and corresponding high spontaneous and inducible mutation rates. Consensus sequence:

```
IMKKKLLLV DGSSLAFFALPPLTNSAGEPTNAVY GFLKMLIKLIEQE QPHTIAVV
F DAKAKTFRHELYEGYKAGRAP
TPDELREQIPLIKELLDALGIPLLEVAGYEADDVIGTLAKLAEKEGYEVLIVTGDRDLL
QLVSDHVTVIITKKGIAEFTL
FTPEAVIEKYGLTPEQIIDYKALMGDSSDNIPGVKGIGEKTA AKLLQEYGSLEGIYANL
DKLKGKKLREKLLAHKEDAKL
SRDLATIKTDVPLDLTLDDLRLPDPDRDALDLLFDE
```

Ref:

Fiorentini P. et al. RT. Mol. Cell. Biol. 17:2764-2773(1997).

Tishkoff et al. Cancer Res. 0:0-0(1998).

Macinnes M.A. et al. Mol. Cell. Biol. 13:6393-6402(1993).

Table A

Pfam	Prosite	Full Name	Description
3_5_exonuclease		3'-5' exonuclease	<p>Accession number: PF01612</p> <p>Definition: 3'-5' exonuclease</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_659 (release 4.1)</p> <p>Gathering cutoffs: -11 -11</p> <p>Trusted cutoffs: -10.70 -10.70</p> <p>Noise cutoffs: -24.50 -24.50</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 85137890</p> <p>Reference Title: Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP.</p> <p>Reference Author: Ollis DL, Brick P, Hamlin R, Xuong NG, Steitz TA;</p> <p>Reference Location: Nature 1985;313:762-766.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98060913</p> <p>Reference Title: The proofreading domain of Escherichia coli DNA polymerase</p> <p>Reference Title: I and other DNA and/or RNA exonuclease domains.</p> <p>Reference Author: Moser MJ, Holley WR, Chatterjee A, Mian IS;</p> <p>Reference Location: Nucleic Acids Res 1997;25:5110-5118.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 98361165</p> <p>Reference Title: Replication focus-forming activity 1 and the Werner syndrome gene product</p> <p>Reference Title: Yan H, Chen CY, Kobayashi R, Newport J;</p> <p>Reference Location: Nat Genet 1998;19:375-378.</p> <p>Reference Number: [4]</p> <p>Reference Medline: 97434221</p> <p>Reference Title: The Werner syndrome protein is a DNA helicase.</p> <p>Reference Author: Gray MD, Shen JC, Kamath-Loeb AS, Blank A, Sopher BL,</p> <p>Reference Author: Martin GM, Oshima J, Loeb LA;</p> <p>Reference Location: Nat Genet 1997;17:100-103.</p> <p>Reference Number: [5]</p> <p>Reference Medline: 97370026</p> <p>Reference Title: DNA helicase activity in Werner's syndrome gene product synthesized in a baculovirus system.</p> <p>Reference Author: Suzuki N, Shimamoto A, Imamura O, Kuromitsu J, Kitao S,</p> <p>Reference Author: Goto M, Furuichi Y;</p> <p>Reference Location: Nucleic Acids Res 1997;25:2973-2978.</p> <p>Database Reference: SCOP; 1dpi; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002562;</p> <p>Database Reference: PDB; 1kfd A; 348; 518;</p> <p>Database Reference: PDB; 1d8y A; 348; 518;</p> <p>Database Reference: PDB; 1d9d A; 348; 518;</p> <p>Database Reference: PDB; 1d9f A; 348; 518;</p> <p>Database Reference: PDB; 1kfs A; 348; 518;</p> <p>Database Reference: PDB; 1kln A; 348; 518;</p> <p>Database Reference: PDB; 1krp A; 348; 518;</p> <p>Database Reference: PDB; 1ksp A; 348; 518;</p> <p>Database Reference: PDB; 1qsl A; 348; 518;</p> <p>Database Reference: PDB; 2kfn A; 348; 518;</p> <p>Database Reference: PDB; 2kfz A; 348; 518;</p> <p>Database Reference: PDB; 2kzm A; 348; 518;</p> <p>Database Reference: PDB; 2kzz A; 348; 518;</p> <p>Comment: This domain is responsible for the 3'-5' exonuclease proofreading</p> <p>Comment: activity of E. coli DNA polymerase I (poll) and other enzymes,</p> <p>Comment: it catalyses the hydrolysis of unpaired or mismatched nucleotides.</p> <p>Comment: This domain consists of the amino-terminal half of the Klenow fragment</p> <p>Comment: in E. coli poll it is also found in the Werner syndrome helicase</p>

			<p>Comment: (WRN), focus forming activity 1 protein (FFA-1) and ribonuclease D</p> <p>Comment: (RNase D).</p> <p>Comment: Werner syndrome is a human genetic disorder causing premature aging;</p> <p>Comment: the WRN protein has helicase activity in the 3'-5' direction [4,5].</p> <p>Comment: The FFA-1 protein is required for formation of a replication foci</p> <p>Comment: and also has helicase activity; it is a homologue of the WRN</p> <p>Comment: protein [3].</p> <p>Comment: RNase D is a 3'-5' exonuclease involved in tRNA processing.</p> <p>Comment: Also found in this family is the autoantigen PM/Scl thought to be</p> <p>Comment: involved in polymyositis-scleroderma overlap syndrome.</p> <p>Number of members: 41</p>
3HCDH	PDOC00065	3-hydroxyacyl-CoA dehydrogenase signature	<p>3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35) (HCDH) [1] is an enzyme involved in fatty acid metabolism, it catalyzes the reduction of 3-hydroxyacyl-CoA to 3-oxoacyl-CoA. Most eukaryotic cells have 2 fatty-acid beta-oxidation systems, one located in mitochondria and the other in peroxisomes. In peroxisomes 3-hydroxyacyl-CoA dehydrogenase forms, with enoyl-CoA hydratase (ECH) and 3,2-trans-enoyl-CoA isomerase (ECI) a multifunctional enzyme where the N-terminal domain bears the hydratase/isomerase activities and the C-terminal domain the dehydrogenase activity. There are two mitochondrial enzymes: one which is monofunctional and the other which is, like its peroxisomal counterpart, multifunctional.</p> <p>In <i>Escherichia coli</i> (gene <i>fadB</i>) and <i>Pseudomonas fragi</i> (gene <i>faoA</i>) HCDH is part of a multifunctional enzyme which also contains an ECH/ECI domain as well as a 3-hydroxybutyryl-CoA epimerase domain [2].</p> <p>The other proteins structurally related to HCDH are:</p> <ul style="list-style-type: none"> - Bacterial 3-hydroxybutyryl-CoA dehydrogenase (EC 1.1.1.157) which reduces 3-hydroxybutanoyl-CoA to acetoacetyl-CoA [3]. - Eye lens protein lambda-crystallin [4], which is specific to lagomorphes (such as rabbit). <p>There are two major region of similarities in the sequences of proteins of the HCDH family, the first one located in the N-terminal, corresponds to the NAD-binding site, the second one is located in the center of the sequence. We have chosen to derive a signature pattern from this central region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [DNE]-x(2)-[GA]-F-[LIVMFY]-x-[NT]-R-x(3)-[PA]-[LIVMFY](2)-x(5)-[LIVMFYCT]-[LIVMFY]-x(2)-[GV]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1998 / Pattern and text revised.</p> <p>References [1] Birktoff J.J., Holden H.M., Hamlin R., Xuong N.-H., Banaszak L.J. Proc. Natl. Acad. Sci. U.S.A. 84:8262-8266(1987).</p> <p>[2] Nakahigashi K., Inokuchi H. Nucleic Acids Res. 18:4937-4937(1990).</p> <p>[3] Mullany P., Clayton C.L., Pallen M.J., Slone R., Al-Saleh A., Tabaqchali S. FEMS Microbiol. Lett. 124:61-67(1994).</p>

			[4] Mulders J.W.M., Hendriks W., Blankesteyn W.M., Bloemendal H., de Jong W.W. J. Biol. Chem. 263:15462-15466(1988).
4HPPD_C		4-hydroxyphenylpyruvate dioxygenase C terminal domain	Accession number: PF01626 Definition: 4-hydroxyphenylpyruvate dioxygenase C terminal domain Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1116 (release 4.1) Gathering cutoffs: -35 -35 Trusted cutoffs: -25.80 -25.80 Noise cutoffs: -44.90 -44.90 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 93279307 Reference Title: Human 4-hydroxyphenylpyruvate dioxygenase. Primary structure and chromosomal localization of the gene. Reference Author: Ruetschi U, Dellsen A, Sahlin P, Stenman G, Rymo L, Lindstedt S; Reference Location: Eur J Biochem 1993;213:1081-1089. Database Reference: INTERPRO; IPR002887; Comment: 4-Hydroxyphenylpyruvic acid dioxygenase (HPD) is an important enzyme Comment: in tyrosine catabolism in most organisms. A genetic deficiency in Comment: this enzyme in humans and mice leads to hereditary tyrosinemia type 3. Comment: The identity of the C-terminus of the HPD makes this part of the Comment: molecule a candidate for a functional role in the catalytic process Comment: [1]. This region is found as a separate protein Swiss:Q49717 that Comment: is somewhat different from HPD and may have a different but related Comment: protein function (Unpublished observation Bateman A). Number of members: 28
5_3_exonuclease		5'-3' exonuclease domain	The 5'-3' exonuclease domain is one found in bacterial DNA polymerases I and in yeast DNA repair enzymes such as Exonuclease I. Yeast Exo I is involved in mitotic recombination and also includes a domain that interacts with the mismatch repair protein MSH2. The 5'-3' exonuclease domain is also present in XPG DNA repair enzymes in humans and in yeast RAD9 protein. Defects in XPG proteins result in Xeroderma Pigmentosum. Thus defects in 5'-3' exonuclease domain-containing proteins in plants are expected to lead to defects in DNA repair and corresponding high spontaneous and inducible mutation rates. Consensus sequence: IMKKKLLLVGDSSLAFFRAFFALPPLTNSAGEPTNAVYGLKMLIKLIEQEQPTHIA VVFDAKAKTFRHELYEGYKAGRAP TPDELREIQIPLIKELLDALGIPLLVAGYEADDVIGTLAKLAEKEGYEVLIVTGDR DLLQLVSDHVTVIITKKGIAEFTL FTPEAVIEKYGLTPEQIIDYKALMGDSSDNIPGVKGIGEKTAALLQEYGSLEGYIY ANLDKLGKGLREKLLAHKEDAKL SRDLATIKTDVPLDLTLDDLRLPDPDRDALDLLFDE Ref: Fiorentini P. et al. RT. Mol. Cell. Biol. 17:2764-2773(1997). Tishkoff et al. Cancer Res. 0:0-0(1998). Macinnes M.A. et al. Mol. Cell. Biol. 13:6393-6402(1993).
60s_ribosomal		60s Acidic ribosomal protein	Accession number: PF00428 Definition: 60s Acidic ribosomal protein Author: Finn RD Alignment method of seed: Clustalw Source of seed members: Pfam-B_151 (release 1.0) Gathering cutoffs: 17 17 Trusted cutoffs: 17.80 17.80 Noise cutoffs: 9.30 9.30 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM

844

			<p>Reference Number: [1] Reference Medline: 96282699 Reference Title: Proteins P1, P2, and P0, components of the eukaryotic ribosome stalk. New structural and functional aspects. Reference Title: ribosome stalk. New structural and functional aspects. Reference Author: Remacha M, Jimenez-Diaz A, Santos C, Briones E, Zambrano R, Reference Author: Rodriguez Gabriel MA, Guarinos E, Ballesta JP; Reference Location: Biochem Cell Biol 1995;73:959-968. Database Reference: INTERPRO; IPR001813; Database reference: PFAMB; PB002218; Comment: This family includes archaeobacterial L12, eukaryotic P0, P1 and P2. Number of members: 109</p>
6PF2K	PDOC00158	Phosphoglycerate mutase family phosphohistidine signature	<p>Phosphoglycerate mutase (EC 5.4.2.1) (PGAM) and bisphosphoglycerate mutase (EC 5.4.2.4) (BPGM) are structurally related enzymes which catalyze reactions involving the transfer of phospho groups between the three carbon atoms of phosphoglycerate [1,2]. Both enzymes can catalyze three different reactions, although in different proportions:</p> <ul style="list-style-type: none"> - The isomerization of 2-phosphoglycerate (2-PGA) to 3-phosphoglycerate (3-PGA) with 2,3-diphosphoglycerate (2,3-DPG) as the primer of the reaction. - The synthesis of 2,3-DPG from 1,3-DPG with 3-PGA as a primer. - The degradation of 2,3-DPG to 3-PGA (phosphatase EC 3.1.3.13 activity). <p>In mammals, PGAM is a dimeric protein. There are two isoforms of PGAM: the M (muscle) and B (brain) forms. In yeast, PGAM is a tetrameric protein. BPGM is a dimeric protein and is found mainly in erythrocytes where it plays a major role in regulating hemoglobin oxygen affinity as a consequence of controlling 2,3-DPG concentration.</p> <p>The catalytic mechanism of both PGAM and BPGM involves the formation of a phosphohistidine intermediate [3].</p> <p>The bifunctional enzyme 6-phosphofructo-2-kinase / fructose-2,6-bisphosphatase (EC 2.7.1.105 and EC 3.1.3.46) (PF2K) [4] catalyzes both the synthesis and the degradation of fructose-2,6-bisphosphate. PF2K is an important enzyme in the regulation of hepatic carbohydrate metabolism. Like PGAM/BPGM, the fructose-2,6-bisphosphatase reaction involves a phosphohistidine intermediate and the phosphatase domain of PF2K is structurally related to PGAM/BPGM.</p> <p>The bacterial enzyme alpha-ribazole-5'-phosphate phosphatase (gene cobC) which is involved in cobalamin biosynthesis also belongs to this family [5].</p> <p>We built a signature pattern around the phosphohistidine residue.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-x-R-H-G-[EQ]-x(3)-N [H is the phosphohistidine residue] Sequences known to belong to this class detected by the pattern ALL, except for Haemophilus influenzae PGAM. Other sequence(s) detected in SWISS-PROT 2.</p> <p>Note some organisms harbor a form of PGAM independent of 2,3-DPG, this enzyme is not related to the family described above [6]. Last update November 1995 / Text revised. References [1] Le Boulch P., Joulin V., Garel M.-C., Rosa J., Cohen-Solal M. Biochem. Biophys. Res. Commun. 156:874-881(1988).</p> <p>[2] White M.F., Fothergill-Gilmore L.A. FEBS Lett. 229:383-387(1988).</p>

			<p>[3] Rose Z.B. Meth. Enzymol. 87:43-51(1982).</p> <p>[4] Bazan J.F., Fletterick R.J., Pilgis S.J. Proc. Natl. Acad. Sci. U.S.A. 86:9642-9646(1989).</p> <p>[5] O'Toole G.A., Trzebiatowski J.R., Escalante-Semerena J.C. J. Biol. Chem. 269:26503-26511(1994).</p> <p>[6] Grana X., De Lecea L., El-Maghrabi M.R., Urena J.M., Caellas C., Carreras J., Puigdomenech P., Pilgis S.J., Climent F. J. Biol. Chem. 267:12797-12803(1992).</p>
7tm_5		7TM chemorecept or	<p>Accession number: PF01604 Definition: 7TM chemoreceptor Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_942 (release 4.1) Gathering cutoffs: -46 -46 Trusted cutoffs: -44.30 -44.30 Noise cutoffs: -47.80 -47.80 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98248686 Reference Title: Two large families of chemoreceptor genes in the nematodes Reference Title: Caenorhabditis elegans and Caenorhabditis briggsae reveal Reference Title: extensive gene duplication, diversification, movement, and Reference Title: intron loss. Reference Author: Robertson HM; Reference Location: Genome Res 1998;8:449-463. Database Reference: INTERPRO; IPR003003; Comment: This large family of proteins are related to 7tm_1. Comment: They are 7 transmembrane receptors. This family does not Comment: include all known members, as there are problems with Comment: overlapping specificity with 7tm_1. Comment: This family is greatly expanded in the nematode worm C. elegans. Number of members: 180</p>
Aa_trans		Transmembrane amino acid transporter protein	<p>Accession number: PF01490 Definition: Transmembrane amino acid transporter protein Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_419 (release 4.0) Gathering cutoffs: 25 25 Trusted cutoffs: 150.80 150.80 Noise cutoffs: 3.60 3.60 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98007977 Reference Title: Identification and characterization of the vesicular GABA transporter. Reference Title: transporter. Reference Author: McIntire SL, Reimer RJ, Schuske K, Edwards RH, Jorgensen Reference Author: EM; Reference Location: Nature 1997;389:870-876. Database Reference: INTERPRO; IPR002422; Database reference: PFAM; PB020912; Comment: This transmembrane region is found in many amino acid transporters Comment: including UNC-47 and MTR. UNC-47 encodes a vesicular amino butyric acid Comment: (GABA) transporter, (VGAT). UNC-47 is predicted to have</p>

846

			<p>10 transmembrane Comment: domains Swiss:P34579 [1]. MTR is a N system amino acid transporter system Comment: protein involved in methyltryptophan resistance Swiss:P38680. Comment: Other members of this family include proline transporters and amino Comment: acid permeases. Number of members: 50</p>
ABC_tran	PDOC00185	ABC transporters family signature	<p>On the basis of sequence similarities a family of related ATP-binding proteins has been characterized [1 to 5]. These proteins are associated with a variety of distinct biological processes in both prokaryotes and eukaryotes, but a majority of them are involved in active transport of small hydrophilic molecules across the cytoplasmic membrane. All these proteins share a conserved domain of some two hundred amino acid residues, which includes an ATP-binding site. These proteins are collectively known as ABC transporters. Proteins known to belong to this family are listed below (references are only provided for recently determined sequences).</p> <p>In prokaryotes:</p> <ul style="list-style-type: none"> - Active transport systems components: alkylphosphonate uptake(phnC/phnK/ phnL); arabinose (araG); arginine (artP); dipeptide (dciAD;dppD/dppF); ferric enterobactin (fepC); ferrichrome (fhuC); galactoside (mglA); glutamine (glnQ); glycerol-3-phosphate (ugpC); glycine betaine/L-proline (proV); glutamate/aspartate (gltL); histidine (hisP); iron(III) (sfuC), iron(III) dicitrate (fecE); lactose (lacK); leucine/isoleucine/valine (braF/braG;livF/livG); maltose (malK); molybdenum (modC); nickel (nikD/ nikE); oligopeptide (amiE/amiF;oppD/oppF); peptide (sapD/sapF); phosphate (pstB); putrescine (potG); ribose (rbsA); spermidine/putrescine (potA); sulfate (cysA); vitamin B12 (btuD). - Hemolysin/leukotoxin export proteins hlyB, cyaB and lktB. - Colicin V export protein cvaB. - Lactococcal export protein lcnC [6]. - Antibiotic transport proteins nisT (nisin) and spaT (subtilin). - Extracellular proteases B and C export protein prtD. - Alkaline protease secretion protein aprD. - Beta-(1,2)-glucan export proteins chvA and ndvA. - Haemophilus influenzae capsule-polysaccharide export protein bexA. - Cytochrome c biogenesis proteins ccmA (also known as cycV and helA). - Polysialic acid transport protein kpsT. - Cell division associated ftsE protein (function unknown). - Copper processing protein nosF from Pseudomonas stutzeri. - Nodulation protein nodI from Rhizobium (function unknown). - Escherichia coli proteins cydC and cydD. - Subunit A of the ABC excision nuclease (gene uvrA). - Erythromycin resistance protein from Staphylococcus epidermidis (gene msrA). - Tylosin resistance protein from Streptomyces fradiae (gene tlrC) [7]. - Heterocyst differentiation protein (gene hetA) from Anabaena PCC 7120. - Protein P29 from Mycoplasma hyorhinis, a probable component of a high affinity transport system. - yhbG, a putative protein whose gene is linked with ntrA in many bacteria such as Escherichia coli, Klebsiella pneumoniae, Pseudomonas putida, Rhizobium meliloti and Thiobacillus ferrooxidans. - Escherichia coli and related bacteria hypothetical proteins yabJ, yadG, yagC, ybbA, ycjW, yddA, yehX, yejF, yheS, yhiG, yhiH, yjcW, yjjK, yojl, yrbF and ytfR. <p>In eukaryotes:</p> <ul style="list-style-type: none"> - The multidrug transporters (Mdr) (P-glycoprotein), a family of closely related proteins which extrude a wide variety of drugs out of the cell (for a review see [8]). - Cystic fibrosis transmembrane conductance regulator (CFTR), which is most probably involved in the transport of chloride ions. - Antigen peptide transporters 1 (TAP1, PSF1, RING4, HAM-1, mtp1) and 2 (TAP2, PSF2, RING11, HAM-2, mtp2), which are involved in the transport of antigens from the cytoplasm to a membrane-bound compartment for association with MHC class I molecules.

		signature	<ul style="list-style-type: none"> - Haemophilus influenzae bexB, involved in polyribosylribitol phosphate capsule polysaccharide export. - Salmonella typhi vexB, involved in translocation of the Vi polysaccharide. - Neisseria meningitidis ctrC, involved in polyneuraminic acid capsule polysaccharide export. - Rhizobiaceae nodulation protein J (gene nodJ), probably involved in exporting a modified beta-1,4-linked N-acetylglucosamine oligosaccharide. - Streptomyces peucetius drrB, involved in exporting the antibiotics daunorubicin and doxorubicin. - Klebsiella pneumoniae O-antigen exprt system protein rfbA. - Yersinia enterocolitica O-antigen exprt system protein rfbD. - Escherichia coli hypothetical protein yadH. - Escherichia coli hypothetical protein yhhJ. <p>The molecular size of these proteins is around 30 Kd. They are thought to contain six transmembrane regions. They either form homooligomeric channels or associate with another type of transmembrane protein to form heteroligomers. Transport systems in which they participate are energized by an ATP-binding protein that belongs to the ABC transporter family. The designation 'ABC-2' has been proposed [1] for these transport systems.</p> <p>As a signature pattern, we selected a conserved region located in the C-terminal section of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIMST]-x(2)-[LIMW]-x(2)-[LIMCA]-[GSTC]-x-[GSAIV]-x(6)-[LIMGA]-[PGSNQ]-x(9,12)-P-[LIMFT]-x-[HRSY]-x(5)-[RQ]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT 2.</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Reizer J., Reizer A., Saier M.H. Jr. Protein Sci. 1:1326-1332(1992).</p> <p>[2] Vazquez M., Santana O., Quinto C. Mol. Microbiol. 8:369-377(1993).</p>
ABC-3		ABC 3 transport family	<p>Members of this family include receptors that mediate transmembrane signalling. These receptors can bind to a number of factors including: amphiregulin, epidermal growth factor, gp30, heparin-binding egf, insulin, insulin-like growth factor I and II, neuregulins, transforming growth factor-alpha and, and vaccinia virus growth</p> <p>Signal transduction is mediated by catalytic activity of tyrosine kinase, such as ATP + A protein tyrosine = ADP + protein tyrosine phosphate. Typically, such signal transduction have been implicated in metabolic and developmental changes, including cell fate and differentiation. Examples include instruction of follicle cells to follow a dorsal pathway of development rather than the default ventral pathway. may also bind the spitz protein.</p> <p>References describing these family members and their biological activities:</p> <p>Abbot et al., J. Biol. Chem. 267:10759-10763(1992); Araki et al., J. Biol. Chem. 262:16186-16191(1987); Aroian et al., EMBO J. 13:360-366(1994); Aroian et al., Nature 348:693-699(1990); Barbetti et al., Diabetes 41:408-415(1992); Bargmann et al., Nature 319:226-230(1986); Cama et al., J. Biol. Chem. 268:8060-8069(1993); Cama et al., J. Clin. Endocrinol. Metab. 73:894-901(1991); Carrera et al., Hum. Mol. Genet. 2:1437-1441(1993); Clifford et al., Genetics 137:531-550(1994); Cocozza et al., Diabetes 41:521-526(1992); Cooke et al., Biochem. Biophys. Res. Commun. 177:1113-1120(1991); Coussens et al., Science 230:1132-1139(1985); Dickens et al., Biochem. Biophys. Res. Commun. 186:244-250(1992); Ebina et al., Cell 40:747-758(1985); Ebina et al., Proc. Natl. Acad. Sci. U.S.A. 84:704-708(1987); Ehsani et al., Genomics 15:426-429(1993); Elbein et al., Diabetes 42:429-434(1993); Elbein, Diabetes 38:737-743(1989); Fujita-Yamaguchi et al., Protein Seq. Data Anal. 1:3-6(1987); Gullick et al., EMBO J. 11:43-48(1992); Haruta et al., Diabetes 42:1837-1844(1993); Hubbard et al., EMBO J. 16:5572-5581(1997).</p>

			<p>Hubbard et al., Nature 372:746-754(1994); Iwanishi et al., Diabetologia 36:414-422(1993); Kadowaki et al., J. Clin. Invest. 86:254-264(1990); Kadowaki et al., Science 240:787-790(1988); Kim et al., Diabetologia 35:261-266(1992); Klinkhamer et al., EMBO J. 8:2503-2507(1989); Kusari et al., J. Biol. Chem. 266:5260-5267(1991); Lai et al., Neuron 6:691-704(1991); Lax et al., Mol. Cell. Biol. 8:1970-1978(1988); Lebrun et al., J. Biol. Chem. 268:11272-11277(1993); Lee et al., Oncogene 8:3403-3410(1993); Lesokhin et al., Dev. Biol. 205:129-144(1999); Livneh et al., Cell 40:599-607(1985).</p> <p>Longo et al., Proc. Natl. Acad. Sci. U.S.A. 90:60-64(1993); McKeon et al., Mol. Endocrinol. 4:647-656(1990); Moller et al., J. Biol. Chem. 265:14979-14985(1990); Moller et al., Mol. Endocrinol. 4:1183-1191(1990); Odawara et al., Science 245:66-68(1989); Raz et al., Genetics 129:191-201(1991).</p> <p>Sakai et al., J. Mol. Biol. 256:548-555(1996); Schaeffer et al., Biochem. Biophys. Res. Commun. 189:650-653(1992); Schejter et al., Cell 46:1091-1101(1986); Seino et al., Biochem. Biophys. Res. Commun. 159:312-316(1989); Seino et al., Diabetes 39:123-128(1990); Semba et al., Proc. Natl. Acad. Sci. U.S.A. 82:6497-6501(1985); Shier et al., J. Biol. Chem. 264:14605-14608(1989); Taira et al., Science 245:63-66(1989); Tewari et al., J. Biol. Chem. 264:16238-16245(1989); Ullrich et al., Nature 313:756-761(1985).</p> <p>Ullrich et al., EMBO J. 5:2503-2512(1986); van der Vorm et al., Diabetologia 36:172-174(1993); van der Vorm et al., J. Biol. Chem. 267:66-71(1992); Wadsworth et al., Nature 314:178-180(1985); White et al., Cell 54:641-649(1988); Xu et al., J. Biol. Chem. 265:18673-18681(1990); Yamamoto et al., Nature 319:230-234(1986); and Yoshimasa et al., Science 240:784-787(1988).</p>
ACAT		Sterol O-acyltransferase	<p>Accession number: PF01800</p> <p>Definition: Sterol O-acyltransferase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1454 (release 4.2)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 112.80 112.80</p> <p>Noise cutoffs: -128.10 -128.10</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98434592</p> <p>Reference Title: Characterization of two human genes encoding acyl coenzyme</p> <p>Reference Title: A:cholesterol acyltransferase-related enzymes.</p> <p>Reference Author: Oelkers P, Behari A, Cromley D, Billheimer JT, Sturley SL;</p> <p>Reference Location: J Biol Chem 1998;273:26765-26771.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98434590</p> <p>Reference Title: Identification of a form of acyl-CoA:cholesterol acyltransferase specific to liver and intestine in nonhuman primates.</p> <p>Reference Title:</p> <p>Reference Author: Anderson RA, Joyce C, Davis M, Reagan JW, Clark M, Shelness</p> <p>Reference Author: GS, Rudel LL;</p> <p>Reference Location: J Biol Chem 1998;273:26747-26754.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 96243137</p> <p>Reference Title: Sterol esterification in yeast: a two-gene process.</p> <p>Reference Author: Yang H, Bard M, Bruner DA, Gleeson A, Deckelbaum RJ,</p> <p>Reference Author: Aljinovic G, Pohl TM, Rothstein R, Sturley SL;</p> <p>Reference Location: Science 1996;272:1353-1356.</p> <p>Database Reference: INTERPRO; IPR002688;</p> <p>Comment: Sterol O-acyltransferases or acyl-coa:cholesterol acyltransferase</p> <p>Comment: (ACAT) EC:2.3.1.26 is a transmembrane protein that catalyses the</p> <p>Comment: esterification of cholesterol to its cholesterol ester storage form.</p> <p>Comment:</p> <p>Number of members: 21</p>
ACPS		4'-phosphopantetheinyl transferase superfamily	<p>Accession number: PF01648</p> <p>Definition: 4'-phosphopantetheinyl transferase superfamily</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1679 (release 4.1)</p>

850

			<p>Gathering cutoffs: 0 0 Trusted cutoffs: 0.60 0.60 Noise cutoffs: -4.00 -4.00 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmlcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96027548 Reference Title: Cloning, overproduction, and characterization of the Reference Title: Escherichia coli holo-acyl carrier protein synthase. Reference Author: Lambalot RH, Walsh CT; Reference Location: J Biol Chem 1995;270:24658-24661. Reference Number: [2] Reference Medline: 97144264 Reference Title: A new enzyme superfamily - the phosphopantetheinyl Reference Title: transferases. Reference Author: Lambalot RH, Gehring AM, Flugel RS, Zuber P, LaCelle M, Reference Author: Marahiel MA, Reid R, Khosla C, Walsh CT; Reference Location: Chem Biol 1996;3:923-936. Reference Number: [3] Reference Medline: 10581256 Reference Title: Crystal structure of the surfactin synthetase-activating Reference Title: enzyme sfp: a prototype of the 4'-phosphopantetheinyl Reference Title: transferase superfamily [In Process Citation] Reference Author: Reuter K, Mofid MR, Marahiel MA, Ficner R; Reference Location: EMBO J 1999;18:6823-6831. Database Reference: INTERPRO; IPR002582; Database reference: PFAMB; PB007908; Database reference: PFAMB; PB041384; Comment: Members of this family transfers the Comment: 4'-phosphopantetheine (4'-PP) moiety from coenzyme A (CoA) to Comment: the invariant serine of pp-binding. This post-translational Comment: modification renders holo-ACP capable of acyl group activation Comment: via thioesterification of the cysteamine thiol of 4'-PP [1]. Comment: This superfamily consists of two subtypes: The ACPs type Comment: such as Swiss:P24224 and the Sfp type such as Swiss:P39135. Comment: The structure of the Sfp type is known [3], which shows the Comment: active site accommodates a magnesium ion. The most highly Comment: conserved regions of the alignment are involved in binding Comment: the magnesium ion. Number of members: 46</p>
ACT		ACT domain	<p>Accession number: PF01842 Definition: ACT domain Author: Bateman A Alignment method of seed: Manual Source of seed members: Bateman A Gathering cutoffs: 25 0 Trusted cutoffs: 26.10 0.50 Noise cutoffs: 24.50 24.50 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmlcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 95236205 Reference Title: The allosteric ligand site in the Vmax-type cooperative Reference Title: enzyme phosphoglycerate dehydrogenase. Reference Author: Schuller DJ, Grant GA, Banaszak LJ; Reference Location: Nat Struct Biol 1995;2:69-76. Reference Number: [2] Reference Medline: 99241053 Reference Title: Gleaning non-trivial structural, functional and Reference Title: evolutionary information about proteins by iterative Reference Title: database searches. Reference Author: Aravind L, Koonin EV; Reference Location: J Mol Biol 1999;287:1023-1040. Database Reference: SCOP; 1psd; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002912; Database Reference: PDB; 1phz A; 35; 110; Database Reference: PDB; 2phm A; 35; 110;</p>

851

			<p>Database Reference PDB; 1psd A; 338; 410; Database Reference PDB; 1psd B; 338; 410; Database reference: PFAMB; PB001977; Database reference: PFAMB; PB008097; Database reference: PFAMB; PB010480; Database reference: PFAMB; PB011031; Database reference: PFAMB; PB031880; Database reference: PFAMB; PB038464; Database reference: PFAMB; PB040963; Database reference: PFAMB; PB041518; Database reference: PFAMB; PB041667; Comment: This family of domains generally have a regulatory role. Comment: ACT domains are linked to a wide range of metabolic Comment: enzymes that are regulated by amino acid concentration. Comment: Pairs of ACT domains bind specifically to a particular Comment: amino acid leading to regulation of the linked enzyme. Comment: The ACT domain is found in: Comment: D-3-phosphoglycerate dehydrogenase EC:1.1.1.95 Swiss:P08328, Comment: which is inhibited by serine [1]. Comment: Aspartokinase EC:2.7.2.4 Swiss:P53553, which is regulated by lysine. Comment: Acetolactate synthase small regulatory subunit Swiss:P00894, Comment: which is inhibited by valine. Comment: Phenylalanine-4-hydroxylase EC:1.14.16.1 Swiss:P00439, which Comment: is regulated by phenylalanine. Comment: Prephenate dehydrogenase EC:4.2.1.51 Swiss:P21203. Comment: formyltetrahydrofolate deformylase EC:3.5.1.10, Swiss:P37051, Comment: which is activated by methionine and inhibited by glycine. Comment: GTP pyrophosphokinase EC:2.7.6.5 Swiss:P11585. Number of members: 177</p>
Acyl-ACP_TE		Acyl-ACP thioesterase	<p>Accession number: PF01643 Definition: Acyl-ACP thioesterase Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_928 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 91.70 91.70 Noise cutoffs: -192.80 -192.80 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96068671 Reference Title: Modification of the substrate specificity of an acyl-acyl carrier protein thioesterase by protein engineering. Reference Author: Yuan L, Voelker TA, Hawkins DJ; Reference Location: Proc Natl Acad Sci U S A 1995;92:10639-10643. Reference Number: [2] Reference Medline: 92320297 Reference Title: Fatty acid biosynthesis redirected to medium chains in transgenic oilseed plants. Reference Title: transgenic oilseed plants. Reference Author: Voelker TA, Worrell AC, Anderson L, Bleibaum J, Fan C, Reference Author: Hawkins DJ, Radke SE, Davies HM; Reference Location: Science 1992;257:72-74. Database Reference INTERPRO; IPR002864; Comment: This family consists of various acyl-acyl carrier protein (ACP) Comment: thioesterases (TE) these terminate fatty acyl group extension via Comment: hydrolyzing an acyl group on a fatty acid [1]. Number of members: 30</p>
Acyltransferase		Acyltransferase	<p>Accession number: PF01553 Definition: Acyltransferase Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_128 (release 4.0) Gathering cutoffs: 8 8 Trusted cutoffs: 14.40 14.40</p>

852

			<p>Noise cutoffs: 2.50 2.50 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 97411131 Reference Title: Barth syndrome may be due to an acyltransferase deficiency. Reference Author: Neuwald AF; Reference Location: Curr Biol 1997;7:465-466. Reference Number: [2] Reference Medline: 96224398 Reference Title: A novel X-linked gene, G4.5. is responsible for Barth syndrome. Reference Author: Bione S, D'Adamo P, Maestrini E, Gedeon AK, Bolhuis PA, Reference Author: Toniolo D; Reference Location: Nat Genet 1996;12:385-389. Database Reference: INTERPRO; IPR002123; Database reference: PFAMB; PB009622; Database reference: PFAMB; PB009717; Database reference: PFAMB; PB033259; Database reference: PFAMB; PB041102; Database reference: PFAMB; PB041638; Comment: This family contains acyltransferases involved in phospholipid biosynthesis and other proteins of unknown function [1]. This family also includes tafazzin Swiss:Q16635, the Barth syndrome gene [2]. Number of members: 74</p>
Adaptin_N		Adaptin N terminal region	<p>Accession number: PF01602 Definition: Adaptin N terminal region Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_491 (release 4.0) Gathering cutoffs: 12 12 Trusted cutoffs: 15.50 15.50 Noise cutoffs: 9.00 9.00 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 97409270 Reference Title: Linking cargo to vesicle formation: receptor tail interactions with coat proteins. Reference Author: Kirchhausen T, Bonifacino JS, Riezman H; Reference Location: Curr Opin Cell Biol 1997;9:488-495. Reference Number: [2] Reference Medline: 89202379 Reference Title: Structural and functional division into two domains of the large (100- to 115-kDa) chains of the clathrin-associated protein complex AP-2. Reference Title: RAKirchhausen T, Nathanson KL, Matsui W, Vaisberg A, Chow Reference Author: EP, Burne C, Keen JH, Davis AE; Reference Location: Proc Natl Acad Sci U S A 1989;86:2612-2616. Database Reference: INTERPRO; IPR002553; Database reference: PFAMB; PB040953; Comment: This family consists of the N terminal region of various alpha, beta and gamma subunits of the AP-1, AP-2 and AP-3 adaptor protein complexes. The adaptor protein (AP) complexes are involved in the formation of clathrin-coated pits and vesicles [1]. Comment: The N-terminal region of the various adaptor proteins (APs) is constant by comparison to the C-terminal which is variable within members of the AP-2 family [2]; and it has been proposed that this constant region interacts with another uniform component of the coated vesicles [2].</p>

853

			Number of members: 66
ALAD	PDOC00153	Delta-aminolevulinic acid dehydratase active site	<p>Delta-aminolevulinic acid dehydratase (EC 4.2.1.24) (ALAD) [1] catalyzes the second step in the biosynthesis of heme, the condensation of two molecules of 5-aminolevulinate to form protobilinogen. The enzyme is an oligomer composed of eight identical subunits. Each of the subunits binds an atom of zinc or of magnesium (in plants). A lysine has been implicated in the catalytic mechanism [2]. The sequence of the region in the vicinity of the active site residue is conserved in ALAD from various prokaryotic and eukaryotic species.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-x-D-x-[LIVM](2)-[IV]-K-P-[GSA]-x(2)-Y [K is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1995 / Pattern and text revised.</p> <p>References [1] Li J.-M., Russell C.S., Cosloy S.D. Gene 75:177-184(1989).</p> <p>[2] Gibbs P.N.B., Jordan P.M. Biochem. J. 236:447-451(1986).</p>
Aldolase	PDOC00144	KDPG and KHG aldolases active site signatures	<p>4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (KHG-aldolase) catalyzes the interconversion of 4-hydroxy-2-oxoglutarate into pyruvate and glyoxylate. Phospho-2-dehydro-3-deoxygluconate aldolase (EC 4.1.2.14) (KDPG-aldolase) catalyzes the interconversion of 6-phospho-2-dehydro-3-deoxy-D-gluconate into pyruvate and glyceraldehyde 3-phosphate.</p> <p>These two enzymes are structurally and functionally related [1]. They are both homotrimeric proteins of approximately 220 amino-acid residues. They are class I aldolases whose catalytic mechanism involves the formation of a Schiff-base intermediate between the substrate and the epsilon-amino group of a lysine residue. In both enzymes, an arginine is required for catalytic activity.</p> <p>We developed two signature patterns for these enzymes. The first one contains the active site arginine and the second, the lysine involved in the Schiff-base formation.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-[LIVM]-x(3)-E-[LIV]-T-[LF]-R [R is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for Bacillus subtilis KDPG-aldolase which has Thr instead of Arg in the active site.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern G-x(3)-[LIVMF]-K-[LF]-F-P-[SA]-x(3)-G [K is involved in Schiff-base formation]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References [1] Vlahos C.J., Dekker E.E. J. Biol. Chem. 263:11683-11691(1988).</p>
Alpha_L_fucosidases	PDOC00324	Alpha-L-fucosidase	<p>Alpha-L-fucosidase (EC 3.2.1.51) [1] is a lysosomal enzyme responsible for hydrolyzing the alpha-1,6-linked fucose joined to the reducing-end N-acetylglucosamine of the carbohydrate moieties of glycoproteins. Deficiency of alpha-L-fucosidase results in the lysosomal storage disease fucosidosis.</p>

854

			<p>A cysteine residue is important for the activity of the enzyme. There is only one cysteine conserved between the sequence of mammalian alpha-L-fucosidase and that of the slime mold Dictyostelium discoideum. We have derived a pattern from the region around that conserved cysteine.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-x(2)-L-x(3)-K-W-E-x-C [C is the putative active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note these proteins belong to family 29 in the classification of glycosyl hydrolases [2,E1].</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Fisher K.J., Aronson N.N. Jr. Biochem. J. 264:695-701(1989).</p> <p>[2] Henrissat B. Biochem. J. 280:309-316(1991).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?glycosid.txt</p>
Amino_oxidase		Flavin containing amine oxidase	<p>Accession number: PF01593</p> <p>Definition: Flavin containing amine oxidase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_606 (release 4.1)</p> <p>Gathering cutoffs: -110 -110</p> <p>Trusted cutoffs: -110.00 -110.00</p> <p>Noise cutoffs: -111.80 -111.80</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98258926</p> <p>Reference Title: Maize polyamine oxidase: primary structure from protein and cDNA sequencing.</p> <p>Reference Author: Tavladoraki P, Schinina ME, Cecconi F, Agostino SD, Manera</p> <p>Reference Author: F, Rea G, Mariottini P, Federico R, Angelini R;</p> <p>Reference Location: FEBS Lett 1998;426:62-66.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97306298</p> <p>Reference Title: A key amino acid responsible for substrate selectivity of monoamine oxidase A and B.</p> <p>Reference Author: Tsugeno Y, Ito A;</p> <p>Reference Location: J Biol Chem 1997;272:14033-14036.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 95287865</p> <p>Reference Title: Cloning, sequencing and heterologous expression of the monoamine oxidase gene from Aspergillus niger.</p> <p>Reference Author: Schilling B, Lerch K;</p> <p>Reference Location: Mol Gen Genet 1995;247:430-438.</p> <p>Database Reference: SCOP; 1b37; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002937;</p> <p>Database Reference: PDB; 1b37 A; 14; 455;</p> <p>Database Reference: PDB; 1b5q A; 14; 455;</p> <p>Database Reference: PDB; 1b37 B; 14; 455;</p> <p>Database Reference: PDB; 1b37 C; 14; 455;</p> <p>Database Reference: PDB; 1b5q B; 14; 455;</p> <p>Database Reference: PDB; 1b5q C; 14; 455;</p> <p>Database reference: PFAMB; PB017518;</p> <p>Database reference: PFAMB; PB024839;</p> <p>Database reference: PFAMB; PB040747;</p> <p>Comment: This family consists of various amine oxidases, including</p>

855

			<p>maze polyamine</p> <p>Comment: oxidase (PAO) [1] and various flavin containing monoamine oxidases</p> <p>Comment: (MAO). The aligned region includes the flavin binding site of these</p> <p>Comment: enzymes.</p> <p>Comment: In vertebrates MAO plays an important role regulating the intracellular</p> <p>Comment: levels of amines via there oxidation; these include various neurotransmitters, neurotoxins and trace amines [2]. In lower eukaryotes</p> <p>Comment: such as aspergillus and in bacteria the main role of amine oxidases is</p> <p>Comment: to provide a source of ammonium [3].</p> <p>Comment: PAOs in plants, bacteria and protozoa oxidase spermidine and spermine</p> <p>Comment: to an aminobutyral, diaminopropane and hydrogen peroxide and are</p> <p>Comment: involved in the catabolism of polyamines [1].</p> <p>Comment: Other members of this family include tryptophan 2-monooxygenase,</p> <p>Comment: putrescine oxidase, corticosteroid binding proteins and antibacterial</p> <p>Comment: glycoproteins.</p> <p>Number of members: 58</p>
ANF receptor	PDOC00430	Natriuretic peptides receptors signature	<p>Natriuretic peptides are hormones involved in the regulation of fluid and electrolyte homeostasis. These hormones stimulate the intracellular production of cyclic GMP as a second messenger.</p> <p>Currently, three types of natriuretic peptide receptors are known [1,2]. Two express guanylate cyclase activity: GC-A (or ANP-A) which seems specific to atrial natriuretic peptide (ANP), and GC-B (or ANP-B) which seems to be stimulated more effectively by brain natriuretic peptide (BNP) than by ANP. The third receptor (ANP-C) is probably responsible for the clearance of ANP from the circulation and does not play a role in signal transduction.</p> <p>GC-A and GC-B are plasma membrane-bound proteins that share the following topology: an N-terminal extracellular domain which acts as the ligand binding region, then a transmembrane domain followed by a large cytoplasmic C-terminal region that can be subdivided into two domains: a protein kinase-like domain (see <PDOC00100>) that appears important for proper signalling and a guanylate cyclase catalytic domain (see <PDOC00425>). The topology of ANP-C is different: like GC-A and -B it possesses an extracellular ligand-binding region and a transmembrane domain, but its cytoplasmic domain is very short.</p> <p>We developed a pattern from the ligand-binding region of natriuretic peptide receptors based on a highly conserved region located in the N-terminal part of the domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-P-x-C-x-Y-x-A-A-x-V-x-R-x(3)-H-W</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update May 1991 / First entry.</p> <p>References [1] Garbers D.L. New Biol. 2:499-504(1990).</p> <p>[2] Schulz S., Chinkers M., Garbers D.L. FASEB J. 2:2026-2035(1989).</p>
Apocytochrome_F	PDOC00169	Cytochrome c family heme-binding site	<p>In proteins belonging to cytochrome c family [1], the heme group is covalently attached by thioether bonds to two conserved cysteine residues. The consensus</p>

856

		signature	<p>sequence for this site is Cys-X-X-Cys-His and the histidine residue is one of the two axial ligands of the heme iron. This arrangement is shared by all proteins known to belong to cytochrome c family, which presently includes cytochromes c, c', c1 to c6, c550 to c556, cc3/Hmc, cytochrome f and reaction center cytochrome c.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-{CPWHF}-{CPWR}-C-H-{CFYW}</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for four cytochrome c's which lack the first thioether bond.</p> <p>Other sequence(s) detected in SWISS-PROT 454.</p> <p>Note: some cytochrome c's have more than a single bound heme group c4 has 2, c7 has 3, c3 has 4, the reaction center has 4, and cc3/Hmc has 16 !</p> <p>Last update June 1992 / Text revised.</p> <p>References [1] Mathews F.S. Prog. Biophys. Mol. Biol. 45:1-56(1985).</p>
arf	PDOC00781 PDOC00017 PDOC01020	ADP-ribosylation factors family signature; ATP/GTP-binding site motif A (P-loop); ATP phosphoribosyltransferase signature PROSITE cross-reference(s)	<p>ADP-ribosylation factors (ARF) [1,2,3,4] are 20 Kd GTP-binding proteins involved in protein trafficking. They may modulate vesicle budding and uncoating within the Golgi apparatus. ARF's also act as allosteric activators of cholera toxin ADP-ribosyltransferase activity. They are evolutionary conserved and present in all eukaryotes. At least six forms of ARF are present in mammals and three in budding yeast. The ARF family also includes proteins highly related to ARF's but which lack the cholera toxin cofactor activity, they are collectively known as ARL's (ARF-like).</p> <p>ARD1 is a 64 Kd mammalian protein of unknown biological function that contains an ARF domain at its C-terminal extremity.</p> <p>Proteins from the ARF family are generally included in the RAS 'superfamily' of small GTP-binding proteins [5], but they are only slightly related to the other RAS proteins. They also differ from RAS proteins in that they lack cysteine residues at their C-termini and are therefore not subject to prenylation. The ARFs are N-terminally myristoylated (the ARLs have not yet been shown to be modified in such a fashion).</p> <p>As a signature pattern, we selected a conserved region in the C-terminal part of ARF's and ARL's.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [HRQT]-x-[FYWI]-x-[LIVM]-x(4)-A-x(2)-G-x(2)-[LIVM]-x(2)-[GSA]-[LIVMF]-x-[WK]-[LIVM]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for 4 sequences.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note proteins belonging to this family also contain a copy of the ATP/GTP-binding motif 'A' (P-loop) (see <PDOC00017>).</p> <p>Expert(s) to contact by email Kahn R.A. rkahn@bimcore.emory.edu</p> <p>Last update November 1997 / Pattern and text revised. Cell. Signal. 4:367-399(1993). References [1] Boman A.L., Kahn R.A. Trends Biochem. Sci. 20:147-150(1995).</p> <p>[2] Moss J., Vaughan M.</p> <p>[3]</p>

857

		<p>Moss J., Vaughan M. Prog. Nucleic Acid Res. Mol. Biol. 45:47-65(1993).</p> <p>[4] Amor J.C., Harrison D.H., Kahn R.A., Ringe D. Nature 372:704-708(1994).</p> <p>[5] Valencia A., Chardin P., Wittinghofer A., Sander C. Biochemistry 30:4637-4648(1991).</p> <p>From sequence comparisons and crystallographic data analysis it has been shown [1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP share a number of more or less conserved sequence motifs. The best conserved of these motifs is a glycine-rich region, which typically forms a flexible loop between a beta-strand and an alpha-helix. This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is generally referred to as the 'A' consensus sequence [1] or the 'P-loop' [5].</p> <p>There are numerous ATP- or GTP-binding proteins in which the P-loop is found. We list below a number of protein families for which the relevance of the presence of such motif has been noted:</p> <ul style="list-style-type: none"> - ATP synthase alpha and beta subunits (see <PDOC00137>). - Myosin heavy chains. - Kinesin heavy chains and kinesin-like proteins (see <PDOC00343>). - Dynamins and dynamin-like proteins (see <PDOC00362>). - Guanylate kinase (see <PDOC00670>). - Thymidine kinase (see <PDOC00524>). - Thymidylate kinase (see <PDOC01034>). - Shikimate kinase (see <PDOC00868>). - Nitrogenase iron protein family (nifH/frxC) (see <PDOC00580>). - ATP-binding proteins involved in 'active transport' (ABC transporters) [7] (see <PDOC00185>). - DNA and RNA helicases [8,9,10]. - GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc.). - Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.). - Nuclear protein ran (see <PDOC00859>). - ADP-ribosylation factors family (see <PDOC00781>). - Bacterial dnaA protein (see <PDOC00771>). - Bacterial recA protein (see <PDOC00131>). - Bacterial recF protein (see <PDOC00539>). - Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc.). - DNA mismatch repair proteins mutS family (See <PDOC00388>). - Bacterial type II secretion system protein E (see <PDOC00567>). <p>Not all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this is the case for tubulins or protein kinases. A special mention must be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [AG]-x(4)-G-K-[ST] Sequences known to belong to this class detected by the pattern a majority. Other sequence(s) detected in SWISS-PROT in addition to the proteins listed above, the 'A' motif is also found in a number of other proteins. Most of these proteins probably bind a nucleotide, but others are definitively not ATP- or GTP-binding (as for example chymotrypsin, or human ferritin light chain). Expert(s) to contact by email Koonin E.V. koonin@ncbi.nlm.nih.gov</p> <p>Last update July 1999 / Text revised. References [1]</p>
--	--	--

858

			<p>Walker J.E., Saraste M., Runswick M.J., Gay N.J. EMBO J. 1:945-951(1982).</p> <p>[2] Moller W., Amons R. FEBS Lett. 186:1-7(1985).</p> <p>[3] Fry D.C., Kuby S.A., Mildvan A.S. Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).</p> <p>[4] Dever T.E., Glynias M.J., Merrick W.C. Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).</p> <p>[5] Saraste M., Sibbald P.R., Wittinghofer A. Trends Biochem. Sci. 15:430-434(1990).</p> <p>[6] Koonin E.V. J. Mol. Biol. 229:1165-1174(1993).</p> <p>[7] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J. Bioenerg. Biomembr. 22:571-592(1990).</p> <p>[8] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).</p> <p>[9] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P.P. Nature 337:121-122(1989).</p> <p>[10] Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M. Nucleic Acids Res. 17:4713-4730(1989).</p> <p>ATP phosphoribosyltransferase (EC 2.4.2.17) is the enzyme that catalyzes the first step in the biosynthesis of histidine in bacteria, fungi and plants. It is a protein of about 23 to 32 Kd. As a signature pattern we selected a region located in the C-terminal part of this enzyme.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern E-x(5)-G-x-[SAG]-x(2)-[IV]-x-D-[LIV]-x(2)-[ST]-G-x-T-[LM] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update July 1998 / First entry.</p>
ArgJ		ArgJ family	<p>Accession number: PF01960 Definition: ArgJ family Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 258.70 99.60 Noise cutoffs: 7.10 7.10 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 93232760 Reference Title: Primary structure, partial purification and regulation of key enzymes of the acetyl cycle of arginine biosynthesis in Reference Title: Bacillus stearothermophilus: dual function of ornithine acetyltransferase. Reference Author: Sakanyan V, Charlier D, Legrain C, Kochikyan A, Mett I, Reference Author: Pierard A, Glansdorff N; Reference Location: J Gen Microbiol 1993;139:393-402. Database Reference INTERPRO; IPR002813;</p>

			<p>Comment: Members of the ArgJ family catalyse the first EC:2.3.1.35 and</p> <p>Comment: fifth steps EC:2.3.1.1 in arginine biosynthesis.</p> <p>Number of members: 22</p>
Armadillo_seg		Armadillo/beta-catenin-like repeats	<p>Accession number: PF00514</p> <p>Definition: Armadillo/beta-catenin-like repeats</p> <p>Author: Bateman A, Chris Ponting, Joerg Schultz, Peer Bork</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: SMART</p> <p>Gathering cutoffs: 24 0</p> <p>Trusted cutoffs: 24.10 0.00</p> <p>Noise cutoffs: 20.70 20.20</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97442350</p> <p>Reference Title: Three-dimensional structure of the armadillo repeat region of beta-catenin.</p> <p>Reference Author: Huber AH, Nelson WJ, Weis WI;</p> <p>Reference Location: Cell 1997;90:871-882.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 96107551</p> <p>Reference Title: Signal transduction of beta-catenin.</p> <p>Reference Author: Gumbiner BM;</p> <p>Reference Location: Curr Opin Cell Biol 1995;7:634-640.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 97454713</p> <p>Reference Title: Armadillo and dTCF: a marriage made in the nucleus.</p> <p>Reference Author: Cavallo R, Rubenstein D, Peifer M;</p> <p>Reference Location: Curr Opin Genet Dev 1997;7:459-466.</p> <p>Reference Number: [4]</p> <p>Reference Medline: 94082295</p> <p>Reference Title: Association of the APC tumor suppressor protein with catenins.</p> <p>Reference Author: Su LK, Vogelstein B, Kinzler KW;</p> <p>Reference Location: Science 1993;262:1734-1737.</p> <p>Reference Number: [5]</p> <p>Reference Medline: 94082294</p> <p>Reference Title: Association of the APC gene product with beta-catenin.</p> <p>Reference Author: Rubinfeld B, Souza B, Albert I, Muller O, Chamberlain SH,</p> <p>Reference Author: Masiarz FR, Munemitsu S, Polakis P;</p> <p>Reference Location: Science 1993;262:1731-1734.</p> <p>Reference Number: [6]</p> <p>Reference Medline: 91084846</p> <p>Reference Title: The segment polarity gene armadillo encodes a functionally</p> <p>Reference Title: modular protein that is the Drosophila homolog of human plakoglobin.</p> <p>Reference Author: Peifer M, Wieschaus E;</p> <p>Reference Location: Cell 1990;63:1167-1176.</p> <p>Database Reference: SCOP; 3bct; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: EXPERT; Chris.Ponting@human-anatomy.oxford.ac.uk;</p> <p>Database reference: SMART; ARM;</p> <p>Database Reference: INTERPRO; IPR000225;</p> <p>Database Reference: PDB; 1ee5 A; 417; 457;</p> <p>Database Reference: PDB; 1bk5 A; 417; 457;</p> <p>Database Reference: PDB; 1bk5 B; 417; 457;</p> <p>Database Reference: PDB; 1bk6 A; 417; 457;</p> <p>Database Reference: PDB; 1bk6 B; 417; 457;</p> <p>Database Reference: PDB; 1ee4 A; 417; 457;</p> <p>Database Reference: PDB; 1ee4 B; 417; 457;</p> <p>Database Reference: PDB; 1ejl I; 409; 449;</p> <p>Database Reference: PDB; 1ejy I; 409; 449;</p> <p>Database Reference: PDB; 1ial A; 409; 449;</p> <p>Database Reference: PDB; 1ee5 A; 246; 286;</p> <p>Database Reference: PDB; 1bk5 A; 246; 286;</p> <p>Database Reference: PDB; 1bk5 B; 246; 286;</p> <p>Database Reference: PDB; 1bk6 A; 246; 286;</p> <p>Database Reference: PDB; 1bk6 B; 246; 286;</p> <p>Database Reference: PDB; 1ee4 A; 246; 286;</p> <p>Database Reference: PDB; 1ee4 B; 246; 286;</p>

		Database Reference	PDB; 1ejl I; 241; 280;
		Database Reference	PDB; 1ejy I; 241; 280;
		Database Reference	PDB; 1ial A; 241; 280;
		Database Reference	PDB; 1ee5 A; 288; 328;
		Database Reference	PDB; 1bk5 A; 288; 328;
		Database Reference	PDB; 1bk5 B; 288; 328;
		Database Reference	PDB; 1bk6 A; 288; 328;
		Database Reference	PDB; 1bk6 B; 288; 328;
		Database Reference	PDB; 1ee4 A; 288; 328;
		Database Reference	PDB; 1ee4 B; 288; 328;
		Database Reference	PDB; 1ejl I; 282; 322;
		Database Reference	PDB; 1ejy I; 282; 322;
		Database Reference	PDB; 1ial A; 282; 322;
		Database Reference	PDB; 1ejl I; 151; 191;
		Database Reference	PDB; 1ejy I; 151; 191;
		Database Reference	PDB; 1ial A; 151; 191;
		Database Reference	PDB; 1ee5 A; 162; 202;
		Database Reference	PDB; 1bk5 A; 162; 202;
		Database Reference	PDB; 1bk5 B; 162; 202;
		Database Reference	PDB; 1bk6 A; 162; 202;
		Database Reference	PDB; 1bk6 B; 162; 202;
		Database Reference	PDB; 1ee4 A; 162; 202;
		Database Reference	PDB; 1ee4 B; 162; 202;
		Database Reference	PDB; 1ee5 A; 330; 370;
		Database Reference	PDB; 1bk5 A; 330; 370;
		Database Reference	PDB; 1bk5 B; 330; 370;
		Database Reference	PDB; 1bk6 A; 330; 370;
		Database Reference	PDB; 1bk6 B; 330; 370;
		Database Reference	PDB; 1ee4 A; 330; 370;
		Database Reference	PDB; 1ee4 B; 330; 370;
		Database Reference	PDB; 1ejl I; 324; 364;
		Database Reference	PDB; 1ejy I; 324; 364;
		Database Reference	PDB; 1ial A; 324; 364;
		Database Reference	PDB; 1ee5 A; 372; 412;
		Database Reference	PDB; 1bk5 A; 372; 412;
		Database Reference	PDB; 1bk5 B; 372; 412;
		Database Reference	PDB; 1bk6 A; 372; 412;
		Database Reference	PDB; 1bk6 B; 372; 412;
		Database Reference	PDB; 1ee4 A; 372; 412;
		Database Reference	PDB; 1ee4 B; 372; 412;
		Database Reference	PDB; 1ejl I; 366; 406;
		Database Reference	PDB; 1ejy I; 366; 406;
		Database Reference	PDB; 1ial A; 366; 406;
		Database Reference	PDB; 1ejl I; 108; 149;
		Database Reference	PDB; 1ejy I; 108; 149;
		Database Reference	PDB; 1ial A; 108; 149;
		Database Reference	PDB; 1ee5 A; 119; 160;
		Database Reference	PDB; 1bk5 A; 119; 160;
		Database Reference	PDB; 1bk5 B; 119; 160;
		Database Reference	PDB; 1bk6 A; 119; 160;
		Database Reference	PDB; 1bk6 B; 119; 160;
		Database Reference	PDB; 1ee4 A; 119; 160;
		Database Reference	PDB; 1ee4 B; 119; 160;
		Database Reference	PDB; 3bct ; 583; 623;
		Database Reference	PDB; 2bct ; 583; 623;
		Database Reference	PDB; 3bct ; 391; 429;
		Database Reference	PDB; 2bct ; 391; 429;
		Database Reference	PDB; 3bct ; 224; 264;
		Database Reference	PDB; 2bct ; 224; 264;
		Database Reference	PDB; 3bct ; 431; 473;
		Database Reference	PDB; 2bct ; 431; 473;
		Database Reference	PDB; 3bct ; 350; 390;
		Database Reference	PDB; 2bct ; 350; 390;
		Database Reference	PDB; 1ejl I; 193; 238;
		Database Reference	PDB; 1ejy I; 193; 238;
		Database Reference	PDB; 1ial A; 193; 238;
		Database Reference	PDB; 1ee5 A; 204; 244;
		Database Reference	PDB; 1bk5 A; 204; 244;
		Database Reference	PDB; 1bk5 B; 204; 244;
		Database Reference	PDB; 1bk6 A; 204; 244;
		Database Reference	PDB; 1bk6 B; 204; 244;
		Database Reference	PDB; 1ee4 A; 204; 244;
		Database Reference	PDB; 1ee4 B; 204; 244;
		Database Reference	PDB; 1ibr D; 399; 437;

861

			<p>Database Reference PDB; 1ibr B; 399; 437; Database Reference PDB; 1qgk A; 399; 437; Database Reference PDB; 1qgr A; 399; 437; Database reference: PFAMB; PB002221; Database reference: PFAMB; PB002617; Database reference: PFAMB; PB004638; Database reference: PFAMB; PB012310; Database reference: PFAMB; PB040528; Database reference: PFAMB; PB041028; Comment: Approx. 40 amino acid repeat. Tandem repeats form super-helix of helices Comment: that is proposed to mediate interaction of beta-catenin with its ligands. Comment: CAUTION: This family does not contain all known armadillo repeats. Number of members: 597</p>
ATP_synt B c	PDOC00137	ATP synthase alpha and beta subunits signature	<p>ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), and a catalytic core, called coupling factor CF(1). The former acts as a proton channel; the latter is composed of five subunits, alpha, beta, gamma, delta and epsilon. The sequences of subunits alpha and beta are related and both contain a nucleotide-binding site for ATP and ADP. The beta chain has catalytic activity, while the alpha chain is a regulatory subunit.</p> <p>Vacuolar ATPases [3] (V-ATPases) are responsible for acidifying a variety of intracellular compartments in eukaryotic cells. Like F-ATPases, they are oligomeric complexes of a transmembrane and a catalytic sector. The sequence of the largest subunit of the catalytic sector (70 Kd) is related to that of F-ATPase beta subunit, while a 60 Kd subunit, from the same sector, is related to the F-ATPases alpha subunit [4].</p> <p>Archaeobacterial membrane-associated ATPases are composed of three subunits. The alpha chain is related to F-ATPases beta chain and the beta chain is related to F-ATPases alpha chain [4].</p> <p>A protein highly similar to F-ATPase beta subunits is found [5] in some bacterial apparatus involved in a specialized protein export pathway that proceeds without signal peptide cleavage. This protein is known as flil in <i>Bacillus</i> and <i>Salmonella</i>, Spa47 (mxlB) in <i>Shigella flexneri</i>, HrpB6 in <i>Xanthomonas campestris</i> and yscN in <i>Yersinia</i> virulence plasmids.</p> <p>In order to detect these ATPase subunits, we took a segment of ten amino-acid residues, containing two conserved serines, as a signature pattern. The first serine seems to be important for catalysis - in the ATPase alpha chain at least - as its mutagenesis causes catalytic impairment.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-[SAP]-[LIV]-[DNH]-x(3)-S-x-S [The first S is a putative active site residue] Sequences known to belong to this class detected by the pattern ALL, except for the archaeobacterium <i>Sulfolobus acidocaldarius</i> ATPase alpha chain where the first Ser is replaced by Gly. Other sequence(s) detected in SWISS-PROT 37.</p> <p>Note F-ATPase alpha and beta subunits, V-ATPase 70 Kd subunit and the archaeobacterial ATPase alpha subunit also contain a copy of the ATP-binding motifs A and B (see <PDOC00017>).</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).</p>

			<p>[2] Senior A.E. Physiol. Rev. 68:177-231(1988).</p> <p>[3] Nelson N. J. Bioenerg. Biomembr. 21:553-571(1989).</p> <p>[4] Gogarten J.P., Kibak H., Dittrich P., Taiz L., Bowman E.J., Bowman B.J., Manolson M.F., Poole R.J., Date T., Oshima T., Konishi J., Denda K., Yoshida M. Proc. Natl. Acad. Sci. U.S.A. 86:6661-6665(1989).</p> <p>[5] Dreyfus G., Williams A.W., Kawagishi I., MacNab R.M. J. Bacteriol. 175:3131-3138(1993).</p>
ATP-synt_D		ATP synthase subunit D	<p>Accession number: PF01813 Definition: ATP synthase subunit D Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1304 (release 4.2) Gathering cutoffs: 25 25 Trusted cutoffs: 157.80 157.80 Noise cutoffs: -79.90 -79.90 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96324968 Reference Title: Subunit structure and organization of the genes of the A1A0 Reference Title: ATPase from the Archaeon Methanosarcina mazei Go1. Reference Author: Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Reference Location: J Biol Chem 1996;271:18843-18852. Reference Number: [2] Reference Medline: 95132627 Reference Title: A bovine cDNA and a yeast gene (VMA8) encoding the subunit Reference Title: D of the vacuolar H(+)-ATPase. Reference Author: Nelson H, Mandiyan S, Nelson N; Reference Location: Proc Natl Acad Sci U S A 1995;92:497-501. Database Reference INTERPRO; IPR002699; Comment: This is a family of subunit D form various ATP synthases Comment: including V-type H+ transporting and Na+ dependent. Comment: Subunit D is suggested to be an integral part of the Comment: catalytic sector of the V-ATPase [2]. Number of members: 21</p>
B56		Protein phosphatase 2A regulatory B subunit (B56 family)	<p>Accession number: PF01603 Definition: Protein phosphatase 2A regulatory B subunit (B56 family) Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_984 (release 4.1) Gathering cutoffs: 11 11 Trusted cutoffs: 17.80 17.80 Noise cutoffs: 5.50 5.50 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96064678 Reference Title: Identification of a new family of protein phosphatase 2A regulatory subunits. Reference Title: regulatory subunits. Reference Author: McCright B, Virshup DM; Reference Location: J Biol Chem 1995;270:26123-26128. Database Reference INTERPRO; IPR002554; Comment: Protein phosphatase 2A (PP2A) is a major intracellular Comment: protein Comment: phosphatase that regulates multiple aspects of cell growth Comment: and metabolism. Comment: The ability of this widely distributed heterotrimeric enzyme to act on a</p>

863

			<p>Comment: diverse array of substrates is largely controlled by the nature of its</p> <p>Comment: regulatory B subunit. There are multiple families of B subunits (See also</p> <p>Comment: PR55), this family is called the B56 family [1].</p> <p>Number of members: 34</p>
Bac_export_1		Bacterial export proteins, family 1	<p>Accession number: PF01311</p> <p>Definition: Bacterial export proteins, family 1</p> <p>Author: Finn RD, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1442 (release 3.0)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 37.20 37.20</p> <p>Noise cutoffs: -95.00 -95.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmbuild --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95113771</p> <p>Reference Title: Caulobacter FliQ and FliR membrane proteins, required for flagellar biogenesis and cell division, belong to a family of virulence factor export proteins.</p> <p>Reference Author: Zhuang WY, Shapiro L;</p> <p>Reference Location: J Bacteriol 1995;177:343-356.</p> <p>Database Reference: INTERPRO; IPR002010;</p> <p>Comment: This family includes the following members;</p> <p>Comment: FliR, MopE, SsaT, YopT, Hrp, HrcT and SpaR</p> <p>Comment: All of these members export proteins, that do not possess signal</p> <p>Comment: peptides, through the membrane. Although the proteins that these</p> <p>Comment: exporters move may be different, the exporters are thought to</p> <p>Comment: function in similar ways [1].</p> <p>Number of members: 29</p>
Band_41	PDOC00566	Band 4.1 family domain signatures and profile	<p>A number of cytoskeletal-associated proteins that associate with various proteins at the interface between the plasma membrane and the cytoskeleton contain a conserved N-terminal domain of about 150 amino-acid residues [1,2,3]. The proteins in which such a domain is known to exist are listed below.</p> <ul style="list-style-type: none"> - Band 4.1, which links the spectrin-actin cytoskeleton of erythrocytes to the plasma membrane. Band 4.1 binds with a high affinity to glycophorin and with lower affinity to band 3 protein. - Ezrin (cytovillin or p81), a component of the undercoat of the microvilli plasma membrane. - Moesin, which is probably involved in binding major cytoskeletal structures to the plasma membrane. - Radixin, which seems to play a crucial role in the binding of the barbed end of actin filaments to the plasma membrane in the undercoat of the cell-to-cell adherens junction (AJ). - Talin, which binds with high affinity to vinculin and with low affinity to integrins. Talin is a high molecular weight (270 Kd) cytoskeletal protein concentrated in regions of cell-substratum contact and, in lymphocytes, of cell-cell contacts. - Filopodin, a slime mold protein that binds actin and which is involved in the control of cell motility and chemotaxis. - Merlin (or schwannomin). Defects in this protein are the cause of type 2 neurofibromatosis (NF2), a predisposition to tumors of the nervous system. - Protein NBL4. - Protein-tyrosine phosphatases PTPN3 (PTP-H1) and PTPN4 (PTP-MEG1). Structurally these two very similar enzymes are composed of a N-terminal band 4.1-like domain followed by a central segment of unknown function and a C-terminal catalytic domain (see <PDOC00323>). They could act at junctions between the membrane and the cytoskeleton. - Protein-tyrosine phosphatases PTPN14 (PEZ or PTP36) and PTP-D1, PTP-RL10 and PTP2E. These phosphatases also consist of a N-terminal band 4.1-like domain and a C-terminal catalytic domain. The central domain seems to contain a SH3-binding domain. - Caenorhabditis elegans protein phosphatase ptp-1.

			<p>Ezrin, moesin, and radixin are highly related proteins, but the other proteins in which this domain is found do not share any region of similarity outside of the domain. In band 4.1 this domain is known to be important for the interaction with glycophorin, an integral membrane protein.</p> <p>We have developed two signature patterns for this domain, one is based on the conserved positions found at the N-terminal extremity of the domain, the second is located in the C-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern W-[LIV]-x(3)-[KRQ]-x-[LIVM]-x(2)-[QH]-x(0,2)-[LIVMF]-x(6,8)-[LIVMF]-x(3,5)-F-[FY]-x(2)-[DENS] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [HYW]-x(9)-[DENQSTV]-[SA]-x(3)-[FY]-[LIVM]-x(2)-[ACV]-x(2)-[LM]-x(2)-[FY]-G-x-[DENQST]-[LIVMFYS] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-PROT 7.</p> <p>Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so. Expert(s) to contact by email Rees J. jrees@vax.oxford.ac.uk</p> <p>Last update November 1997 / Patterns and text revised; profile added.</p> <p>References [1] Rees D.J.G., Ades S.A., Singer S.J., Hynes R.O. Nature 347:685-689(1990).</p> <p>[2] Funayama N., Nagafuchi A., Sato N., Tsukita S., Tsukita S. J. Cell Biol. 115:1039-1048(1991).</p> <p>[3] Takeuchi K., Kawashima A., Nagafuchi A., Tsukita S. J. Cell Sci. 107:1921-1928(1994).</p>
biotin_lipoyl	PDOC00167; PDOC00168	Biotin-requiring enzymes; 2-oxo acid dehydrogenases acyltransferases component lipoyl binding	<p>Biotin, which plays a catalytic role in some carboxyl transfer reactions, is covalently attached, via an amide bond, to a lysine residue in enzymes requiring this coenzyme [1,2,3,4]. Such enzymes are:</p> <ul style="list-style-type: none"> - Pyruvate carboxylase (EC 6.4.1.1). - Acetyl-CoA carboxylase (EC 6.4.1.2). - Propionyl-CoA carboxylase (EC 6.4.1.3). - Methylcrotonoyl-CoA carboxylase (EC 6.4.1.4). - Geranoyl-CoA carboxylase (EC 6.4.1.5). - Urea carboxylase (EC 6.3.4.6). - Oxaloacetate decarboxylase (EC 4.1.1.3). - Methylmalonyl-CoA decarboxylase (EC 4.1.1.41). - Glutaconyl-CoA decarboxylase (EC 4.1.1.70). - Methylmalonyl-CoA carboxyl-transferase (EC 2.1.3.1) (transcarboxylase). <p>Sequence data reveal that the region around the biocytin (biotin-lysine) residue is well conserved and can be used as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GN]-[DEQTR]-x-[LIVMFY]-x(2)-[LIVM]-x-[AIV]-M-K-[LMAT]-x(3)-[LIVM]-x-[SAV] [K is the biotin attachment site] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p>

			<p>Note the domain around the biotin-binding lysine residue is evolutionary related to that around the lipoyl-binding lysine residue of 2-oxo acid dehydrogenase acyltransferases (see <PDOC00168>).</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References</p> <p>[1] Knowles J.R. Annu. Rev. Biochem. 58:195-221(1989).</p> <p>[2] Samols D., Thronton C.G., Murtif V.L., Kumar G.K., Haase F.C., Wood H.G. J. Biol. Chem. 263:6461-6464(1988).</p> <p>[3] Goss N.H., Wood H.G. Meth. Enzymol. 107:261-278(1984).</p> <p>[4] Shenoy B.C., Xie Y., Park V.L., Kumar G.K., Beegen H., Wood H.G., Samols D. J. Biol. Chem. 267:18407-18412(1992).</p> <p>The 2-oxo acid dehydrogenase multienzyme complexes [1,2] from bacterial and eukaryotic sources catalyze the oxidative decarboxylation of 2-oxo acids to the corresponding acyl-CoA. The three members of this family of multienzyme complexes are:</p> <ul style="list-style-type: none"> - Pyruvate dehydrogenase complex (PDC). - 2-oxoglutarate dehydrogenase complex (OGDC). - Branched-chain 2-oxo acid dehydrogenase complex (BCOADC). <p>These three complexes share a common architecture: they are composed of multiple copies of three component enzymes - E1, E2 and E3. E1 is a thiamine pyrophosphate-dependent 2-oxo acid dehydrogenase, E2 a dihydrolipamide acyltransferase, and E3 an FAD-containing dihydrolipamide dehydrogenase.</p> <p>E2 acyltransferases have an essential cofactor, lipoic acid, which is covalently bound via a amide linkage to a lysine group. The E2 components of OGDC and BCOADC bind a single lipoyl group, while those of PDC bind either one (in yeast and in Bacillus), two (in mammals), or three (in Azotobacter and in Escherichia coli) lipoyl groups [3].</p> <p>In addition to the E2 components of the three enzymatic complexes described above, a lipoic acid cofactor is also found in the following proteins:</p> <ul style="list-style-type: none"> - H-protein of the glycine cleavage system (GCS) [4]. GCS is a multienzyme complex of four protein components, which catalyzes the degradation of glycine. H protein shuttles the methylamine group of glycine from the P protein to the T protein. H-protein from either prokaryotes or eukaryotes binds a single lipoic group. - Mammalian and yeast pyruvate dehydrogenase complexes differ from that of other sources, in that they contain, in small amounts, a protein of unknown function - designated protein X or component X. Its sequence is closely related to that of E2 subunits and seems to bind a lipoic group [5]. - Fast migrating protein (FMP) (gene acoC) from Alcaligenes eutrophus [6]. This protein is most probably a dihydrolipamide acyltransferase involved in acetoin metabolism. <p>We developed a signature pattern which allows the detection of the lipoyl-binding site.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GN]-x(2)-[LIVF]-x(5)-[LIVFC]-x(2)-[LIVFA]-x(3)-K-[STAIV]-[STAVQDN]-x(2)-[LIVMFS]-x(5)-[GCN]-x-[LIVMFY] [K is the lipoyl-binding site] Sequences known to belong to this class detected by the pattern ALL.</p>
--	--	--	--

866

			<p>Other sequence(s) detected in SWISS-PROT 2.</p> <p>Note the domain around the lipoyl-binding lysine residue is evolutionary related to that around the biotin-binding lysine residue of biotin requiring enzymes (see <PDOC00167>).</p> <p>Last update November 1995 / Text revised.</p> <p>References</p> <p>[1] Yeaman S.J. Biochem. J. 257:625-632(1989).</p> <p>[2] Yeaman S.J. Trends Biochem. Sci. 11:293-296(1986).</p> <p>[3] Russel G.C., Guest J.R. Biochim. Biophys. Acta 1076:225-232(1991).</p> <p>[4] Fujiwara K., Okamura-Ikeda K., Motokawa Y. J. Biol. Chem. 261:8836-8841(1986).</p> <p>[5] Behal R.H., Browning K.S., Hall T.B., Reed L.J. Proc. Natl. Acad. Sci. U.S.A. 86:8732-8736(1989).</p> <p>[6] Priefert H., Hein S., Krueger N., Zeh K., Schmidt B., Steinbuechel A. J. Bacteriol. 173:4056-4071(1991).</p>
Biotin_synth		Biotin synthase	<p>Accession number: PF01792</p> <p>Definition: Biotin synthase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1407 (release 4.2)</p> <p>Gathering cutoffs: -180 -180</p> <p>Trusted cutoffs: -176.30 -176.30</p> <p>Noise cutoffs: -183.90 -183.90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96312354</p> <p>Reference Title: Cloning, sequencing, and characterization of the Bacillus subtilis biotin biosynthetic operon.</p> <p>Reference Author: Bower S, Perkins JB, Yocum RR, Howitt CL, Rahaim P, Pero J;</p> <p>Reference Location: J Bacteriol 1996;178:4122-4130.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97074643</p> <p>Reference Title: Two new members of the bio B superfamily: cloning, sequencing and expression of bio B genes of Methylobacillus</p> <p>Reference Title: flagellatum and Corynebacterium glutamicum.</p> <p>Reference Author: Serebriiskii IG, Vassin VM, Tsygankov YD;</p> <p>Reference Location: Gene 1996;175:15-22.</p> <p>Database Reference: INTERPRO: IPR002684;</p> <p>Database reference: PFAMB; PB023954;</p> <p>Database reference: PFAMB; PB040740;</p> <p>Database reference: PFAMB; PB041208;</p> <p>Comment: Biotin synthase EC:2.8.1.6 works with flavodoxin, S-adenosylmethionine,</p> <p>Comment: and possibly cysteine to convert dethiobiotin to biotin [1].</p> <p>Comment: Biotin (vitamin H) is a prosthetic group in enzymes catalysing</p> <p>Comment: carboxylation and transcarboxylation reactions [2].</p> <p>Number of members: 29</p>
BolA		BolA-like protein	<p>Accession number: PF01722</p> <p>Definition: BolA-like protein</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1996 (release 4.1)</p>

867

			<p>Gathering cutoffs: 23 23 Trusted cutoffs: 23.70 23.70 Noise cutoffs: -16.00 -16.00 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 99291046 Reference Title: The stationary-phase morphogene bolA from <i>Escherichia coli</i> Reference Title: is induced by stress during early stages of growth. Reference Author: Santos JM, Freire P, Vicente M, Arraiano CM; Reference Location: Mol Microbiol 1999;32:789-798. Reference Number: [2] Reference Medline: 90059998 Reference Title: Induction of a growth-phase-dependent promoter triggers transcription of bolA, an <i>Escherichia coli</i> morphogene. Reference Author: Aldea M, Garrido T, Hernandez-Chico C, Vicente M, Kushner Reference Author: SR; Reference Location: EMBO J 1989;8:3923-3931. Database Reference: INTERPRO; IPR002634; Comment: This family consist of the morpho-protein BolA from Comment: <i>E. coli</i> and its various homologs. In <i>E. coli</i> over expression of Comment: this protein causes round morphology and may be involved in Comment: switching the cell between elongation and septation Comment: systems during Comment: cell division [1]. The expression of BolA is growth rate regulated Comment: and is induced during the transition into the the stationary phase [1]. BolA is also induced by stress during early stages of Comment: growth [1] and may have a general role in stress response. Comment: It has also been suggested that BolA can induce the transcription Comment: of penicillin binding proteins 6 and 5 [2,1]. Number of members: 18</p>
casein_kappa			<p>Accession number: PF00997 Definition: Kappa casein Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1298 (release 3.0) Gathering cutoffs: -32 -32 Trusted cutoffs: 16.40 16.40 Noise cutoffs: -73.00 -73.00 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98072500 Reference Title: Nucleotide sequence evolution at the kappa-casein locus: Reference Title: evidence for positive selection within the family Bovidae. Reference Author: Ward TJ, Honeycutt RL, Derr JN; Reference Location: Genetics 1997;147:1863-1872. Database Reference: INTERPRO; IPR000117; Comment: Kappa-casein is a mammalian milk protein involved in a number of important physiological processes. In the gut, Comment: the ingested protein is split into an insoluble peptide (para kappa-casein) and a soluble hydrophilic Comment: glycopeptide Comment: (caseinomacropptide). Caseinomacropptide is responsible Comment: for increased efficiency of digestion, prevention of neonate Comment: hypersensitivity to ingested proteins, and inhibition of Comment: gastric pathogens. Number of members: 56</p>
CAT	PDOC00093	Chloramphenicol acetyltransferase	<p>Chloramphenicol acetyltransferase (CAT) (EC 2.3.1.28) [1] catalyzes the acetyl-CoA dependent acetylation of chloramphenicol (Cm), an antibiotic which inhibits prokaryotic peptidyltransferase activity. Acetylation of Cm by CAT inactivates the antibiotic. A histidine residue, located in the C-terminal section of the enzyme, plays a central role in its catalytic mechanism. We</p>

868

			<p>derived a signature pattern from the region surrounding this active site residue.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern Q-[LIV]-H-H-[SA]-x(2)-D-G-[FY]-H [The second H is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note there is a second family of CAT [2], evolutionary unrelated to the main family described above. These CAT belong to the bacterial hexapeptide-repeat containing-transferases family (see <PDOC00094>).</p> <p>Last update November 1997 / Text revised.</p> <p>References [1] Shaw W.V., Leslie A.G.W. Annu. Rev. Biophys. Chem. 20:363-386(1991).</p> <p>[2] Parent R., Roy P.H. J. Bacteriol. 174:2891-2897(1992).</p>
Cation_efflux		Cation efflux family	<p>Accession number: PF01545 Definition: Cation efflux family Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_232 (release 4.0) Gathering cutoffs: -6 -6 Trusted cutoffs: 6.90 6.90 Noise cutoffs: -19.30 -19.30 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98361887 Reference Title: Molecular characterization of a chromosomal determinant conferring resistance to zinc and cobalt ions in Staphylococcus aureus. Reference Author: Xiong A, Jayaswal RK; Reference Location: J Bacteriol 1998;180:4024-4029. Reference Number: [2] Reference Medline: 96219090 Reference Title: Cloning and sequence analysis of czc genes in Alcaligenes sp. strain CT14. Reference Author: Kunito T, Kusano T, Oyaizu H, Senoo K, Kanazawa S, Matsumoto S; Reference Location: Biosci Biotechnol Biochem 1996;60:699-704. Database Reference: INTERPRO; IPR002524; Database reference: PFAM; PB038216; Comment: Members of this family are integral membrane proteins, that Comment: are found to increase tolerance to divalent metal ions such as cadmium, zinc, and cobalt. These proteins are thought to Comment: be efflux pumps that remove these ions from cells. Number of members: 59</p>
CBD_6		Cellulose binding domain	<p>Accession number: PF02018 Definition: Cellulose binding domain Author: Bateman A Alignment method of seed: Manual Source of seed members: Chris Ponting Gathering cutoffs: 19 0 Trusted cutoffs: 19.10 19.10 Noise cutoffs: 8.90 8.90 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97074498 Reference Title: Structure of the N-terminal cellulose-binding domain of</p>

			<p>Reference Title: Cellulomonas fimi CenC determined by nuclear magnetic resonance spectroscopy.</p> <p>Reference Author: Johnson PE, Joshi MD, Tomme P, Kilburn DG, McIntosh LP;</p> <p>Reference Location: Biochemistry 1996;35:14381-14394.</p> <p>Database Reference: URL; http://www.ocms.ox.ac.uk/~ponting/methmb/example.html;</p> <p>Database Reference: SCOP; 1ulp; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: PDB; 1ulo ; 1; 149;</p> <p>Database Reference: PDB; 1ulp ; 1; 149;</p> <p>Database Reference: PDB; 1cx1 A; 2; 6;</p> <p>Database Reference: PDB; 1ulo ; 150; 152;</p> <p>Database Reference: PDB; 1ulp ; 150; 152;</p> <p>Database Reference: PDB; 1cx1 A; 7; 151;</p> <p>Database reference: PFAMB; PB012497;</p> <p>Database reference: PFAMB; PB041237;</p> <p>Database reference: PFAMB; PB041605;</p> <p>Number of members: 76</p>
CBFD_NFYB_HMF	PDOC00578	CBF/NF-Y subunits signatures	<p>Diverse DNA binding proteins are known to bind the CCAAT box, a common cis-acting element found in the promoter and enhancer regions of a large number of genes in eukaryotes. Amongst these proteins is one known as the CCAAT-binding factor (CBF) or NF-Y [1]. CBF is a heteromeric transcription factor that consists of two different components both needed for DNA-binding.</p> <p>The HAP protein complex of yeast binds to the upstream activation site of cytochrome C iso-1 gene (CYC1) as well as other genes involved in mitochondrial electron transport and activates their expression. It also recognizes the sequence CCAAT and is structurally and evolutionary related to CBF.</p> <p>The first subunit of CBF, known as CBF-A or NF-YB in vertebrates, HAP3 in budding yeast and as php3 in fission yeast, is a protein of 116 to 210 amino-acid residues which contains a highly conserved central domain of about 90 residues. This domain seems to be involved in DNA-binding; we have developed a signature pattern from its central part.</p> <p>The second subunit of CBF, known as CBF-B or NF-YA in vertebrates, HAP2 in budding yeast and php2 in fission yeast, is a protein of 265 to 350 amino-acid residues which contains a highly conserved region of about 60 residues. This region, called the 'essential core' [2], seems to consist of two subdomains: an N-terminal subunit-association domain and a C-terminal DNA recognition domain. We have developed a signature pattern from a section of the subunit-association domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-V-S-E-x-I-S-F-[LIVM]-T-[SG]-E-A-[SC]-[DE]-[KRQ]-C Sequences known to belong to this class detected by the pattern ALL CBF-A subunits. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern Y-V-N-A-K-Q-Y-x-R-I-L-K-R-R-x-A-R-A-K-L-E Sequences known to belong to this class detected by the pattern ALL CBF-B subunits. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1995 / Patterns and text revised.</p> <p>References [1] Li X.-Y., Mantovani R., Hooft van Huijsduijnen R., Andre I., Benoist C., Mathis D. Nucleic Acids Res. 20:1087-1091(1992).</p> <p>[2] Olesen J.T., Fikes J.D., Guarente L. Mol. Cell. Biol. 11:611-619(1991).</p>

870

CbiX		CbiX	<p>Accession number: PF01903 Definition: CbiX Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: -25 -25 Trusted cutoffs: -23.10 -23.10 Noise cutoffs: -35.10 -35.10 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98416126 Reference Title: Cobalamin (vitamin B12) biosynthesis: identification and Reference Title: characterization of a <i>Bacillus megaterium</i> cobl operon. Reference Author: Raux E, Lanois A, Warren MJ, Rambach A, Thermes C; Reference Location: Biochem J 1998;335:159-166. Database Reference: INTERPRO; IPR002762; Database reference: PFAMB; PB040604; Database reference: PFAMB; PB040610; Database reference: PFAMB; PB041575; Comment: The function of CbiX is uncertain, however it is found Comment: in cobalamin biosynthesis operons and so may have a Comment: related function. Some CbiX proteins contain a striking Comment: histidine-rich region at their C-terminus, which suggests Comment: that it might be involved in metal chelation [1]. Number of members: 6</p>
cellulase	PDOC00565	Glycosyl hydrolases family 5 signature	<p>The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family A [3] or as the glycosyl hydrolases family 5 [4,E1]. The enzymes which are currently known to belong to this family are listed below.</p> <ul style="list-style-type: none"> - Endoglucanases from various species and strains of <i>Bacillus</i>. - <i>Butyrivibrio fibrisolvens</i> endoglucanases 1 (end1) and A (celA). - <i>Caldocellum saccharolyticum</i> bifunctional endoglucanase/exoglucanase (celB). - This protein consists of two domains; it is the C-terminal domain, which has endoglucanase activity, which belongs to this family. - <i>Clostridium acetobutylicum</i> endoglucanase (eglA). - <i>Clostridium cellulolyticum</i> endoglucanases A (celcA) and D (celcD). - <i>Clostridium cellulovorans</i> endoglucanase B (engB) and D (engD). - <i>Clostridium thermocellum</i> endoglucanases B (celB), C (celC), E (celE), G (celG) and H (celH). - <i>Erwinia chrysanthemi</i> endoglucanase Z (celZ). - <i>Fibrobacter succinogenes</i> endoglucanase 3 (cel-3). - <i>Pseudomonas fluorescens</i> endoglucanase C (celC). - <i>Pseudomonas solanacearum</i> endoglucanase (egl). - <i>Robillarda</i> strain Y-20 endoglucanase I. - <i>Ruminococcus albus</i> endoglucanases I (EG-I), A (celA), and B (celB). - <i>Ruminococcus flavefaciens</i> cellodextrinase A (celA). - <i>Ruminococcus flavefaciens</i> endoglucanase E (celE). - <i>Streptomyces lividans</i> endoglucanase. - <i>Thermomonospora fusca</i> endoglucanase E-5 (celE). - <i>Trichoderma reesei</i> endoglucanase II (EGLII). - <i>Xanthomonas campestris</i> endoglucanase (engxA). <p>As well as:</p> <ul style="list-style-type: none"> - Baker's yeast glucan 1,3-beta-glucosidase I/II (EC 3.2.1.58) (EXG1). - Baker's yeast glucan 1,3-beta-glucosidase 2 (EC 3.2.1.58) (EXG2). - Baker's yeast sporulation-specific glucan 1,3-beta-glucosidase (SPR1). - <i>Caldocellum saccharolyticum</i> beta-mannanase (EC 3.2.1.78) (manA). - Yeast hypothetical protein YBR056w. - Yeast hypothetical protein YIR007w. <p>One of the conserved regions in these enzymes contains a conserved glutamic acid residue which is potentially involved [5] in the catalytic mechanism. We use this region as a signature pattern.</p>

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIV]-[LIVMFYWGA](2)-[DNEQG]-[LIVMGST]-x-N-E-[PV]-[RHDNSTLIVFY] [E is a putative active site residue] Sequences known to belong to this class detected by the pattern ALL, except for Robillarda Y-20 endoglucanase I whose sequence is known to be incorrect and yeast YBR056w. Other sequence(s) detected in SWISS-PROT 22. Expert(s) to contact by email Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).</p> <p>[2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).</p> <p>[3] Henrissat B., Claeysens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).</p> <p>[4] Henrissat B. Biochem. J. 280:309-316(1991).</p> <p>[5] Py B., Bortoli-German I., Haiech J., Chippaux M., Barras F. Protein Eng. 4:325-333(1991).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?glycosid.txt</p>
CH	PDOC00019	Actinin-type actin-binding domain signatures	<p>Alpha-actinin is a F-actin cross-linking protein which is thought to anchor actin to a variety of intracellular structures [1]. The actin-binding domain of alpha-actinin seems to reside in the first 250 residues of the protein. A similar actin-binding domain has been found in the N-terminal region of many different actin-binding proteins [2,3]:</p> <ul style="list-style-type: none"> - In the beta chain of spectrin (or fodrin). - In dystrophin, the protein defective in Duchenne muscular dystrophy (DMD) and which may play a role in anchoring the cytoskeleton to the plasma membrane. - In the slime mold gelation factor (or ABP-120). - In actin-binding protein ABP-280 (or filamin), a protein that link actin filaments to membrane glycoproteins. - In fimbrin (or plastin), an actin-bundling protein. Fimbrin differs from the above proteins in that it contains two tandem copies of the actin-binding domain and that these copies are located in the C-terminal part of the protein. <p>We selected two conserved regions as signature patterns for this type of domain. The first of this region is located at the beginning of the domain, while the second one is located in the central section and has been shown to be essential for the binding of actin.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [EQ]-x(2)-[ATV]-[FY]-x(2)-W-x-N Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 25.</p> <p>Consensus pattern [LIVM]-x-[SGN]-[LIVM]-[DAGHE]-[SAG]-x-[DNEAG]-[LIVM]-x-[DEAG]-x(4)-[LIVM]-x-[LM]-[SAG]-[LIVM]-[LIVMT]-W-x-[LIVM](2) Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p>

			<p>Last update November 1997 / Patterns and text revised.</p> <p>References</p> <p>[1] Schleicher M., Andre E., Harmann A., Noegel A.A. Dev. Genet. 9:521-530(1988).</p> <p>[2] Matsudaira P. Trends Biochem. Sci. 16:87-92(1991).</p> <p>[3] Dubreuil R.R. BioEssays 13:219-226(1991).</p>
chitinase_2	PDOC00839	Chitinases family 18 active site	<p>Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1]. Chitinases of family 18 (also known as classes III or V) groups a variety of proteins:</p> <p>a) Chitinases from:</p> <ul style="list-style-type: none"> - Prokaryotes such as Alteromonas, Bacillus, Serratia, Streptomyces, etc. - Plants such as Arabidopsis, cucumber, bean, tobacco, etc. - Fungi such as Aphanocladium, Rhizopus, Saccharomyces, etc. - Nematode (Brugia malayi). - Insects (Manduca sexta). - Baculoviruses (Autographa Californica Nuclear Polyhedrosis virus). <p>b) Other proteins:</p> <ul style="list-style-type: none"> - Hevamine, a rubber tree protein with chitinase and lysozyme activities. - Kluyveromyces lactis killer toxin alpha subunit, which acts as a chitinase. - Flavobacterium and Streptomyces endo-beta-N-acetylglucosaminidases (EC 3.2.1.96). - Mammalian di-N-acetylchitobiase which is involved in the degradation of asparagine-linked glycoproteins. - Human cartilage glycoprotein Gp-39. - Jack bean concanavalin B (conB), a protein that has lost its catalytic activity. <p>Site directed mutagenesis experiments [3] and crystallographic data [4,5] have shown that a conserved glutamate is involved in the catalytic mechanism and probably acts as a proton donor. This glutamate is at the extremity of the best conserved region in these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMFY]-[DN]-G-[LIVMF]-[DN]-[LIVMF]-[DN]-x-E [E is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for conB which has a Gln instead of the active site Glu.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Expert(s) to contact by email Neuhaus J.-M. jean-marc.neuhaus@bota.unine.ch</p> <p>Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update November 1997 / Text revised.</p> <p>References</p> <p>[1] Flach J., Pilet P.-E., Jolles P. Experientia 48:701-716(1992).</p> <p>[2] Henrissat B. Biochem. J. 280:309-316(1991).</p> <p>[3]</p>

873

			<p>Watanabe T., Kohori K., Miyashita K., Fujii T., Sakai H., Uchida M., Tanaka H. J. Biol. Chem. 268:18567-18572(1993).</p> <p>[4] Perrakis A., Tews I., Dauter Z., Oppenheim A.B., Chet I., Wilson K.S., Vorgias C.E. Structure 2:1169-1180(1994).</p> <p>[5] van Scheltinga A.C.T., Kalk K.H., Beintema J.J., Dijkstra B.W. Structure 2:1181-1189(1994).</p> <p>[E1] http://www.expasy.ch/cgi-bi/lists?glycosid.txt</p>
Choline_kinase		Choline/ethanolamine kinase	<p>Accession number: PF01633 Definition: Choline/ethanolamine kinase Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1165 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 242.90 242.90 Noise cutoffs: -85.90 -85.90 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmlcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98175949 Reference Title: Expression, purification, and characterization of choline kinase, product of the CKI gene from <i>Saccharomyces cerevisiae</i>. Reference Author: Kim KH, Voelker DR, Flocco MT, Carman GM; Reference Location: J Biol Chem 1998;273:6844-6852. Database Reference: INTERPRO; IPR002573; Comment: Choline kinase catalyses the committed step in the synthesis of Comment: phosphatidylcholine by the CDP-choline pathway [1]. Number of members: 22</p>
Chorion		Chorion protein	<p>Accession number: PF01723 Definition: Chorion protein Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1914 (release 4.1) Gathering cutoffs: -46 -46 Trusted cutoffs: -43.70 -43.70 Noise cutoffs: -49.00 -49.00 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmlcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 95333194 Reference Title: Sequence analysis of a small early chorion gene subfamily Reference Title: interspersed within the late gene locus in <i>Bombyx mori</i>. Reference Author: Kravariti L, Lecanidou R, Rodakis GC; Reference Location: J Mol Evol 1995;41:24-33. Reference Number: [2] Reference Medline: 86313609 Reference Title: Evolution of the silk moth chorion gene superfamily: gene families CA and CB. Reference Author: Lecanidou R, Rodakis GC, Eickbush TH, Kafatos FC; Reference Location: Proc Natl Acad Sci U S A 1986;83:6514-6518. Database Reference: INTERPRO; IPR002635; Database reference: PFAMB; PB009425; Comment: This family consists of the chorion superfamily proteins classes A, B, CA, CB and high-cysteine HCB from silk, gypsy and polyphemus moths. Comment: The chorion proteins make up the moths egg shell a complex Comment: extracellular structure [2]. Number of members: 35</p>
Chorismate_mut		Chorismate mutase	<p>Accession number: PF01817 Definition: Chorismate mutase Author: Bateman A</p>

874

			<p>Alignment method of seed: Manual Source of seed members: PSI-BLAST 1ecm Gathering cutoffs: 5 5 Trusted cutoffs: 5.10 5.10 Noise cutoffs: -19.90 -19.90 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 95062155 Reference Title: The crystal structure of allosteric chorismate mutase at 2.2-A resolution. Reference Author: Xue Y, Lipscomb WN, Graf R, Schnappauf G, Braus G; Reference Location: Proc Natl Acad Sci U S A 1994;91:10814-10818. Reference Number: [2] Reference Medline: 98307941 Reference Title: Tyrosine and tryptophan act through the same binding site Reference Title: at the dimer interface of yeast chorismate mutase. Reference Author: Schnappauf G, Krappmann S, Braus GH; Reference Location: J Biol Chem 1998;273:17012-17017. Reference Number: [3] Reference Medline: 98165805 Reference Title: Chorismate mutase-prephenate dehydratase from Escherichia coli. Study of catalytic and regulatory domains using genetically engineered proteins. Reference Author: Zhang S, Pohnert G, Kongsaree P, Wilson DB, Clardy J, Reference Author: Ganem B; Reference Location: J Biol Chem 1998;273:6248-6253. Database Reference: SCOP; 1csm; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002701; Database Reference: PDB; 1ecm B; 6; 89; Database Reference: PDB; 1ecm A; 5; 89; Database Reference: PDB; 1csm A; 133; 162; Database Reference: PDB; 3csm A; 133; 243; Database Reference: PDB; 3csm B; 133; 243; Database Reference: PDB; 4csm A; 133; 243; Database Reference: PDB; 4csm B; 133; 243; Database Reference: PDB; 5csm A; 133; 243; Database Reference: PDB; 2csm A; 133; 246; Comment: Chorismate mutase EC:5.4.99.5 catalyses the conversion of Comment: chorismate to prephenate in the pathway of tyrosine and phenylalanine biosynthesis. This enzyme is negatively Comment: regulated by tyrosine, tryptophan and phenylalanine [2,3]. Number of members: 28</p>
CN_hydrolase	PDOC00712; PDOC00943	Nitrilases / cyanide hydratase signatures; Uncharacterized protein family UPF0012 signature	<p>Nitrilases (EC 3.5.5.1) are enzymes that convert nitriles into their corresponding acids and ammonia. They are widespread in microbes as well as in plants where they convert indole-3-acetonitrile to the hormone indole-3-acetic acid. A conserved cysteine has been shown [1,2] to be essential for enzyme activity; it seems to be involved in a nucleophilic attack on the nitrile carbon atom.</p> <p>Cyanide hydratase (EC 4.2.1.66) converts HCN to formamide. In phytopathogenic fungi, it is used to avoid the toxic effect of cyanide released by wounded plants [3]. The sequence of cyanide hydrolase is evolutionary related to that of nitrilases.</p> <p>Yeast hypothetical proteins YIL164c and YIL165c also belong to this family.</p> <p>As signature patterns for these enzymes, we selected two conserved regions. The first is located in the N-terminal section while the second, which contains the active site cysteine, is located in the central section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-x(2)-[LIVMFY](2)-x-[IF]-x-E-x(2)-[LIVM]-x-G-Y-P Sequences known to belong to this class detected by the pattern ALL.</p>

875

			<p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern G-[GAQ]-x(2)-C-[WA]-E-[NH]-x(2)-[PST]-[LIVMFYS]-x-[KR] [C is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1995 / Patterns and text revised.</p> <p>References [1] Kobayashi M., Izui H., Nagasawa T., Yamada H. Proc. Natl. Acad. Sci. U.S.A. 90:247-251(1993).</p> <p>[2] Kobayashi M., Komeda H., Yanaka N., Nagasawa T., Yamada H. J. Biol. Chem. 267:20746-20751(1992).</p> <p>[3] Wang P., Vanetten H.D. Biochem. Biophys. Res. Commun. 187:1048-1054(1992).</p> <p>The following uncharacterized proteins have been shown [1] to share regions of similarities:</p> <ul style="list-style-type: none"> - Yeast chromosome X hypothetical protein YJL126w. - Yeast chromosome XII hypothetical protein YLR351c. - Fission yeast hypothetical protein SpAC26A3.11. - Escherichia coli hypothetical protein ybeM. - Bacillus subtilis hypothetical protein yhcX. - Mycobacterium tuberculosis hypothetical protein MtCY20G9.06c. - Synechocystis strain PCC 6803 hypothetical protein sl0601. - A Pseudomonas fluorescens hypothetical protein in pqqF 5' region. - A Staphylococcus hypothetical protein in agr operon. <p>Except for yhcX which is larger, these are protein of about 30 to 35 Kd which contain, in their central section, a well conserved region centered on a cysteine residue.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GTA]-x(2)-[IVT]-C-Y-D-[LIVM]-x-F-P-x(9)-G</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / First entry.</p> <p>References [1] Bairoch A. Unpublished observations (1995).</p>
CorA		CorA-like Mg2+ transporter protein	<p>Accession number: PF01544</p> <p>Definition: CorA-like Mg2+ transporter protein</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_944 (release 4.0)</p> <p>Gathering cutoffs: -62 -62</p> <p>Trusted cutoffs: -5.90 -5.90</p> <p>Noise cutoffs: -86.20 -86.20</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98448512</p> <p>Reference Title: The CorA magnesium transporter gene family.</p> <p>Reference Author: Kehres DG, Lawyer CH, Maguire ME;</p> <p>Reference Location: Microb Comp Genomics 1998;3:151-169.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 99003207</p> <p>Reference Title: The CorA Mg2+ transport protein of Salmonella typhimurium.</p> <p>Reference Title: Mutagenesis of conserved residues in the third membrane</p> <p>Reference Title: domain identifies a Mg2+ pore.</p>

			<p>Reference Author: Smith RL, Szegedy MA, Kucharski LM, Walker C, Wiet RM, Reference Author: Redpath A, Kaczmarek MT, Maguire ME; Reference Location: J Biol Chem 1998;273:28663-28669. Database Reference: INTERPRO; IPR002523; Database reference: PFAMB; PB041399; Comment: The CorA transport system is the primary Mg2+ influx system of Salmonella Comment: typhimurium and Escherichia coli. CorA is virtually ubiquitous in the Comment: Bacteria and Archaea. There are also eukaryotic relatives of this protein Number of members: 25</p>
Cys_knot	PDOC00234	Glycoprotein hormones beta chain signatures	<p>Glycoprotein hormones [1,2] (or gonadotropins) are a family of proteins which include the mammalian hormones follitropin (FSH), lutropin (LSH), thyrotropin (TSH) and chorionic gonadotropin (CG), as well as at least two forms of fish gonadotropins. All these hormones consist of two glycosylated chains (alpha and beta). In mammalian gonadotropins, the alpha chain is identical in the four types of hormones but the beta chains, while homologous, are different.</p> <p>The beta chains are proteins of about 100 to 140 amino acid residues which contain twelve conserved cysteines all involved in disulfide bonds [3], as shown in the following schematic representation.</p> <pre> +-----+ +-----+-----+ +-----+-----+ +-----+-----+ **** ***** xxx Cxxxxxx CxCxCxCxxxxxx Cxxxxxx Cxxxxxx CxCxCxCxxxx Cxxxxxx xxx +-----+ +-----+ </pre> <p>'C': conserved cysteine involved in a disulfide bond. *': position of the patterns.</p> <p>We have developed two patterns for these hormones. The first one, located in the N-terminal section, is a region which has been said to be involved in the association between the two chains of the hormones. The second pattern consists of a cluster of five conserved cysteines in the C-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-[STAGM]-G-[HFYL]-C-x-[ST] [The two C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern ALL, except for rat beta-FSH which has Glu in position 2 of the pattern. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [PA]-V-A-x(2)-C-x-C-x(2)-C-x(4)-[STD]-[DEY]-C-x(6,8)-[PGSTAVM]-x(2)-C [The five C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern ALL, except for 5 sequences. Other sequence(s) detected in SWISS-PROT NONE. Expert(s) to contact by email Laphorn A. adrian@chem.gla.ac.uk</p> <p>Last update July 1998 / Patterns and text revised. References [1] Pierce J.G., Parsons T.F. Annu. Rev. Biochem. 50:465-495(1981). [2] Stockell Hartree A., Renwick A.G.C. Biochem. J. 287:665-679(1992).</p>

			<p>[3] Lapthorn A.J., Harris D.C., Littlejohn A., Lustbader J.W., Canfield R.E., Machin K.J., Morgan F.J., Isaacs N.W. Nature 369:455-461(1994).</p>
cytochrome_b_C	PDOC00171	Cytochrome b/b6 signatures	<p>In the mitochondrion of eukaryotes and in aerobic prokaryotes, cytochrome b is a component of respiratory chain complex III (EC 1.10.2.2) - also known as the bc1 complex or ubiquinol-cytochrome c reductase. In plant chloroplasts and cyanobacteria, there is a analogous protein, cytochrome b6, a component of the plastoquinone-plastocyanin reductase (EC 1.10.99.1), also known as the b6f complex.</p> <p>Cytochrome b/b6 [1,2] is an integral membrane protein of approximately 400 amino acid residues that probably has 8 transmembrane segments. In plants and cyanobacteria, cytochrome b6 consists of two subunits encoded by the petB and petD genes. The sequence of petB is colinear with the N-terminal part of mitochondrial cytochrome b, while petD corresponds to the C-terminal part. Cytochrome b/b6 non-covalently binds two heme groups, known as b562 and b566.</p> <p>Four conserved histidine residues are postulated to be the ligands of the iron atoms of these two heme groups.</p> <p>Apart from regions around some of the histidine heme ligands, there are a few conserved regions in the sequence of b/b6. The best conserved of these regions includes an invariant P-E-W triplet which lies in the loop that separates the fifth and sixth transmembrane segments. It seems to be important for electron transfer at the ubiquinone redox site - called Qz or Qo (where o stands for outside) - located on the outer side of the membrane.</p> <p>A schematic representation of the structure of cytochrome b/b6 is shown below.</p> <pre> +---Fe-b562---+ +---Fe-b566--- + xxxxxxxxxxxHxHxxxxxxxxxxxxHxHxxxxxxxxPEWxxxxxxxxxxxxxxxx <-----Cytochrome-b-----> <---Cytochrome-b6-petB-----> <---Cytochrome-b6-petD-----> </pre> <p>We developed two signature patterns for cytochrome b/b6. The first includes the first conserved histidine of b/b6, which is a heme b562 ligand; the second includes the conserved PEW triplet.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [DENQ]-x(3)-G-[FYWMQ]-x-[LIVMF]-R-x(2)-H [H is a heme b562 ligand] Sequences known to belong to this class detected by the pattern ALL, except for 5 sequences. Other sequence(s) detected in SWISS-PROT 15.</p> <p>Consensus pattern P-[DE]-W-[FY]-[LFY](2) Sequences known to belong to this class detected by the pattern ALL, except for Odocoileus hemionus (mule deer) and Paramecium tetraurelia cytochrome b. Other sequence(s) detected in SWISS-PROT 1.</p> <p>Last update November 1995 / Patterns and text revised.</p> <p>References [1] Howell N. J. Mol. Evol. 29:157-169(1989).</p> <p>[2] Esposti M.D., de Vries S., Crimi M., Ghelli A., Patarnello T., Meyer A. Biochim. Biophys. Acta 1143:243-271(1993).</p>
cytochrome_b_N	PDOC00171	Cytochrome b/b6 signatures	<p>In the mitochondrion of eukaryotes and in aerobic prokaryotes, cytochrome b is a component of respiratory chain complex III (EC 1.10.2.2) - also known as the bc1 complex or ubiquinol-cytochrome c reductase. In plant chloroplasts and cyanobacteria, there is a analogous protein, cytochrome b6, a component of the plastoquinone-plastocyanin reductase (EC 1.10.99.1), also known as the b6f complex.</p>

879

			<p>for four cytochrome c's which lack the first thioether bond. Other sequence(s) detected in SWISS-PROT 454.</p> <p>Note: some cytochrome c's have more than a single bound heme group c4 has 2, c7 has 3, c3 has 4, the reaction center has 4, and cc3/Hmc has 16 !</p> <p>Last update June 1992 / Text revised. References [1] Mathews F.S. Prog. Biophys. Mol. Biol. 45:1-56(1985).</p>
DAHP_synth_2		Class-II DAHP synthetase family	<p>Members of this family are aldolase enzymes that catalyse the first step of the shikimate pathway. These polypeptides can be useful in the synthesis of aromatic compounds, such as amino acids, antibiotics, secondary metabolites, etc. Such synthesis can occur either in vitro or in vivo.</p>
Dala_Dala_lig as		D-ala D-ala ligase	<p>Accession number: PF01820 Definition: D-ala D-ala ligase Author: Bateman A Alignment method of seed: Clustalw Source of seed members: PSI-BLAST 2dln Gathering cutoffs: 25 25 Trusted cutoffs: 44.90 26.60 Noise cutoffs: 21.50 18.90 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97207065 Reference Title: D-alanine:D-alanine ligase: phosphonate and phosphinate Reference Title: intermediates with wild type and the Y216F mutant. Reference Author: Fan C, Park IS, Walsh CT, Knox JR; Reference Location: Biochemistry 1997;36:2531-2538. Database Reference: SCOP; 2dln; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR000291; Database Reference: PDB; 1iov ; 3; 303; Database Reference: PDB; 1iow ; 3; 303; Database Reference: PDB; 2dln ; 3; 303; Comment: This family contains D-alanine--D-alanine ligase enzymes EC:6.3.2.4. Number of members: 80</p>
DHPS	PDOC00630	Dihydropteroate synthase signatures	<p>All organisms require reduced folate cofactors for the synthesis of a variety of metabolites. Most microorganisms must synthesize folate de novo because they lack the active transport system of higher vertebrate cells which allows these organisms to use dietary folates. Enzymes that are involved in the biosynthesis of folates are therefore the target of a variety of antimicrobial agents such as trimethoprim or sulfonamides.</p> <p>Dihydropteroate synthase (EC 2.5.1.15) (DHPS) catalyzes the condensation of 6-hydroxymethyl-7,8-dihydropteridine pyrophosphate to para-aminobenzoic acid to form 7,8-dihydropteroate. This is the second step in the three steps pathway leading from 6-hydroxymethyl-7,8-dihydropterin to 7,8-dihydrofolate. DHPS is the target of sulfonamides which are substrates analog that compete with para-aminobenzoic acid.</p> <p>Bacterial DHPS (gene sul or folP) [1] is a protein of about 275 to 315 amino acid residues which is either chromosomally encoded or found on various antibiotic resistance plasmids. In the lower eukaryote <i>Pneumocystis carinii</i>, DHPS is the C-terminal domain of a multifunctional folate synthesis enzyme (gene fas) [2].</p> <p>We developed two signature patterns for DHPS, the first signature is located in the N-terminal section of these enzymes, while the second signature is located in the central section.</p> <p>Description of pattern(s) and/or profile(s)</p>

880

			<p>Consensus pattern [LIVM]-x-[AG]-[LIVMF](2)-N-x-T-x-D-S-F-x-D-x-[SG] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [GE]-[SA]-x-[LIVM](2)-D-[LIVM]-G-[GP]-x(2)-[STA]-x-P Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References [1] Slock J., Stahly D.P., Han C.-Y., Six E.W., Crawford I.P. J. Bacteriol. 172:7211-7226(1990).</p> <p>[2] Volpes F., Dyer M., Scaife J.G., Darby G., Stammers D.K., Delves C.J. Gene 112:213-218(1992).</p>
DHquinase_I	PDOC00788	Dehydroquinase class I active site	<p>3-dehydroquinase dehydratase (EC 4.2.1.10), or dehydroquinase, catalyzes the conversion of 3-dehydroquinase into 3-dehydroshikimate. It is the third step in the shikimate pathway for the biosynthesis of aromatic amino acids from chorismate. Two classes of dehydroquinases exist, known as types I and II. The best studied type I enzyme is from <i>Escherichia coli</i> (gene <i>aroD</i>) and related bacteria where it is a homodimeric protein of a chain of about 250 residues. In fungi, dehydroquinase is part of a multifunctional enzyme which catalyzes five consecutive steps in the shikimate pathway. In <i>aroD</i>, it has been shown [1] that a histidine is involved in the catalytic mechanism; we used the region around this residue as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern D-[LIVM]-[DE]-[LIVMN]-x(18,20)-[LIVM](2)-x-[SC]-[NHY]-H-[DN] [H is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update December 1999 / Pattern and text revised.</p> <p>References [1] Deka R.K., Kleanthous C., Coggins J.R. J. Biol. Chem. 267:22237-22242(1992).</p>
Diphthamide_syn		Putative diphthamide synthesis protein	<p>Accession number: PF01866 Definition: Putative diphthamide synthesis protein Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 44.90 44.90 Noise cutoffs: -174.70 -174.70 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96183112 Reference Title: A cDNA from the ovarian cancer critical region of deletion on chromosome 17p13.3. Reference Author: Phillips NJ, Zeigler MR, Deaven LL; Reference Location: Cancer Lett 1996;102:85-90. Reference Number: [2] Reference Medline: 94010339 Reference Title: Diphthamide synthesis in <i>Saccharomyces cerevisiae</i>: structure of the DPH2 gene. Reference Author: Mattheakis LC, Sor F, Collier RJ; Reference Location: Gene 1993;132:149-154. Database Reference INTERPRO: IPR002728; Comment: Swiss:Q16439 is a candidate tumour suppressor gene [1]. DPH2 from Comment: yeast Swiss:P32461 [2], which confers resistance to diphtheria toxin Comment: has been found to be involved in diphthamide synthesis. Diphtheria</p>

			<p>Comment: toxin inhibits eukaryotic protein synthesis by ADP-ribosylating</p> <p>Comment: diphthamide, a posttranslationally modified histidine residue present</p> <p>Comment: in EF2. The exact function of the members of this family is</p> <p>Comment: unknown.</p> <p>Number of members: 12</p>
DLH		Dienelactone hydrolase family	<p>Accession number: PF01738</p> <p>Definition: Dienelactone hydrolase family</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_757 (release 4.2)</p> <p>Gathering cutoffs: 15 0</p> <p>Trusted cutoffs: 15.60 3.10</p> <p>Noise cutoffs: 14.40 14.40</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 90339491</p> <p>Reference Title: Refined structure of dienelactone hydrolase at 1.8 A.</p> <p>Reference Author: Pathak D, Ollis D;</p> <p>Reference Location: J Mol Biol 1990;214:497-525.</p> <p>Database Reference: SCOP; 1din; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002925;</p> <p>Database Reference: PDB; 1din ; 16; 232;</p> <p>Database reference: PFAMB; PB004640;</p> <p>Database reference: PFAMB; PB041131;</p> <p>Database reference: PFAMB; PB041469;</p> <p>Number of members: 42</p>
DNA_mis_rep air	PDOC00057	DNA mismatch repair proteins mutL / hexB / PMS1 signature	<p>Mismatch repair contributes to the overall fidelity of DNA replication [1]. It involves the correction of mismatched base pairs that have been missed by the proofreading element of the DNA polymerase complex. The sequence of some proteins involved in mismatch repair in different organisms have been found to be evolutionary related. These proteins are:</p> <ul style="list-style-type: none"> - Escherichia coli and Salmonella typhimurium mutL protein [2]. MutL is required for dam-dependent methyl-directed DNA repair. - Streptococcus pneumoniae hexB protein [3]. The Hex system is nick directed. - Yeast proteins PMS1 and MLH1 [4]. - Human protein MLH1 [5] which is involved in a form of familial hereditary nonpolyposis colon cancer (HNPCC). <p>As a signature pattern for this class of mismatch repair proteins we selected a perfectly conserved heptapeptide which is located in the N-terminal section of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-F-R-G-E-A-L</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update</p> <p>November 1995 / Pattern and text revised.</p> <p>References</p> <p>[1]</p> <p>Modrich P.</p> <p>Annu. Rev. Biochem. 56:435-466(1987).</p> <p>[2]</p> <p>Mankovich J.A., McIntyre C.A., Walker G.C.</p> <p>J. Bacteriol. 171:5325-5331(1989).</p> <p>[3]</p> <p>Prudhomme M., Martin B., Mejean V., Claverys J.-P.</p> <p>J. Bacteriol. 171:5332-5338(1989).</p> <p>[4]</p> <p>Prolla T.A., Christie D., Liskay R.M.</p>

			<p>Mol. Cell. Biol. 14:407-415(1994).</p> <p>[5] Bronner C.E., Baker S.M., Morrison P.T., Warren G., Smith L.G., Lescoe M.K., Kane M., Earibino C., Lipford J., Linblom A., Tannergard P., Bollag R.J., Godwin A.R., Ward D.C., Nordenskjold M., Fishel R., Kolodner R.D., Liskay R.M. Nature 368:258-261(1994).</p>
DNA primase _S		DNA primase small subunit	<p>Accession number: PF01896 Definition: DNA primase small subunit Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 198.40 198.40 Noise cutoffs: -120.80 -120.80 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 91219475 Reference Title: Mutations in conserved yeast DNA primase domains impair DNA Reference Title: replication in vivo. Reference Author: Francesconi S, Longhese MP, Piseri A, Santocanale C, Reference Author: Lucchini G, Plevani P; Reference Location: Proc Natl Acad Sci U S A 1991;88:3877-3881. Database Reference INTERPRO; IPR002755; Comment: DNA primase synthesizes the RNA primers for the Okazaki Comment: fragments in lagging strand DNA synthesis. DNA primase Comment: is a heterodimer of large and small subunits. Number of members: 14</p>
DnaB		DnaB-like helicase	<p>Members of this family are comprise DNA replication enzymes which unwind the helix. Generally, such polypeptide are ATPases which move at the replication fork, disrupting hydrogen bonds. Such proteins are use for DNA replication in vivo and/or in vitro.</p>
DnaJ_C		DnaJ C terminal region	<p>Accession number: PF01556 Definition: DnaJ C terminal region Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_342 (release 4.0) Gathering cutoffs: -24 -24 Trusted cutoffs: -22.60 -22.60 Noise cutoffs: -25.50 -25.50 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98308847 Reference Title: The J-domain family and the recruitment of chaperone power. Reference Author: Kelley WL; Reference Location: Trends Biochem Sci 1998;23:222-227. Database Reference INTERPRO; IPR002939; Database reference: PFAMB; PB013976; Comment: This family consists of the C terminal region form the DnaJ Comment: protein. Although the function of this region is unknown, it Comment: is always found associated with DnaJ and DnaJ_CXXCXGXG. Comment: DnaJ is a chaperone associated with the Hsp70 heat-shock Comment: system involved in protein folding and renaturation after stress. Number of members: 116</p>
DnaJ_CXXCXGXG	PDOC00553	dnaJ domains signatures and profile	<p>The prokaryotic heat shock protein dnaJ interacts with the chaperone hsp70-like dnaK protein [1]. Structurally, the dnaJ protein consists of an N-terminal conserved domain (called 'J' domain) of about 70 amino acids, a glycine-rich region ('G' domain) of about 30 residues, a central domain containing four repeats of a CXXCXGXG motif ('CRR' domain) and a C-terminal region of 120 to 170 residues. Such a structure is shown in the following</p>

883

		<p>schematic representation:</p> <pre> +-----+-----+-----+-----+ N-terminal Gly-R CXXCXGXG C-terminal +-----+-----+-----+-----+ </pre> <p>It has been shown [2] that the 'J' domain as well as the 'CRR' domain are also found in other prokaryotic and eukaryotic proteins which are listed below.</p> <p>a) Proteins containing both a 'J' and a 'CRR' domain:</p> <ul style="list-style-type: none"> - Yeast protein MAS5/YDJ1 which seems to be involved in mitochondrial protein import. - Yeast protein MDJ1, involved in mitochondrial biogenesis and protein folding. - Yeast protein SCJ1, involved in protein sorting. - Yeast protein XDJ1. - Plants dnaJ homologs (from leek and cucumber). - Human HDJ2, a dnaJ homolog of unknown function. - Yeast hypothetical protein YNL077w. <p>b) Proteins containing a 'J' domain without a 'CRR' domain:</p> <ul style="list-style-type: none"> - Rhizobium fredii nolC, a protein involved in cultivar-specific nodulation of soybean. - Escherichia coli cbpA [3], a protein that binds curved DNA. - Yeast protein SEC63/NPL1, important for protein assembly into the endoplasmic reticulum and the nucleus. - Yeast protein SIS1, required for nuclear migration during mitosis. - Yeast protein CAJ1. - Yeast hypothetical protein YFR041c. - Yeast hypothetical protein YIR004w. - Yeast hypothetical protein YJL162c. - Plasmodium falciparum ring-infected erythrocyte surface antigen (RESA). RESA, whose function is not known, is associated with the membrane skeleton of newly invaded erythrocytes. - Human HDJ1. - Human HSJ1, a neuronal protein. - Drosophila cysteine-string protein (csp). <p>We developed a signature pattern for the 'J' domain, based on conserved positions in the C-terminal half of this domain. We also developed a pattern for the 'CRR' domain, based on the first two copies of that motif. We also developed a profile for the 'J' domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [FY]-x(2)-[LIVMA]-x(3)-[FYWHNT]-[DENQSA]-x-L-x-[DN]-x(3)-[KR]-x(2)-[FYI] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 5.</p> <p>Consensus pattern C-[DEGSTHKR]-x-C-x-G-x-[GK]-[AGSDM]-x(2)-[GSNKR]-x(4,6)-C-x(2,3)-C-x-G-x-G Sequences known to belong to this class detected by the pattern ALL, except for yeast XDJ1. Other sequence(s) detected in SWISS-PROT 8.</p> <p>Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.</p> <p>Expert(s) to contact by email Kelley W. kelley@medecine.unige.ch</p> <p>Last update July 1998 / Patterns and text revised. References</p>
--	--	---

			<p>[1] Cyr D.M., Langer T., Douglas M.G. Trends Biochem. Sci. 19:176-181(1994).</p> <p>[2] Bork P., Sander C., Valencia A., Bukau B. Trends Biochem. Sci. 17:129-129(1992).</p> <p>[3] Ueguchi C., Kaneda M., Yamada H., Mizuno T. Proc. Natl. Acad. Sci. U.S.A. 91:1054-1058(1994).</p>
dNK		Deoxynucleo side kinase	<p>Accession number: PF01712 Definition: Deoxynucleoside kinase Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1744 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 47.50 47.50 Noise cutoffs: -5.40 -5.40 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97236800 Reference Title: Cloning of the cDNA and chromosome localization of the gene Reference Title: for human thymidine kinase 2. Reference Author: Johansson M, Karlsson A; Reference Location: J Biol Chem 1997;272:8454-8458. Reference Number: [2] Reference Medline: 96293511 Reference Title: Cloning and expression of human deoxyguanosine kinase cDNA. Reference Author: Johansson M, Karlsson A; Reference Location: Proc Natl Acad Sci U S A 1996;93:7258-7262. Database Reference INTERPRO; IPR002624; Comment: This family consists of various deoxynucleoside kinases Comment: cytidine EC:2.7.1.74, guanosine EC:2.7.1.113, adenosine EC:2.7.1.76 Comment: and thymidine kinase EC:2.7.1.21 (which also phosphorylates deoxyuridine Comment: and deoxycytosine.) These enzymes catalyse the production of Comment: deoxynucleotide 5'-monophosphate from a deoxynucleoside. Comment: Using ATP and yielding ADP in the process. Number of members: 20</p>
DUF125		Integral membrane protein DUF125	<p>Accession number: PF01988 Definition: Integral membrane protein DUF125 Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: -60 -60 Trusted cutoffs: -57.90 -57.90 Noise cutoffs: -64.60 -64.60 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 95028150 Reference Title: Sequence, mapping and disruption of CCC1, a gene that Reference Title: cross-complements the Ca(2+)-sensitive phenotype of csg1 Reference Title: mutants. Reference Author: Fu D, Beeler T, Dunn T; Reference Location: Yeast 1994;10:515-521. Database Reference INTERPRO; IPR002839; Comment: This family of predicted integral membrane proteins has no known Comment: function. However it does include Swiss:P47818, that may have a Comment: role in regulating calcium levels [1]. Number of members: 7</p>

885

DUF25		Domain of unknown function DUF25	<p>Accession number: PF01641</p> <p>Definition: Domain of unknown function DUF25</p> <p>Author: Bateman A, Enwright A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1539 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 151.80 151.80</p> <p>Noise cutoffs: 10.60 10.60</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 20076492</p> <p>Reference Title: Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif.</p> <p>Reference Author: Lescure A, Gautheret D, Carbon P, Krol A;</p> <p>Reference Location: J Biol Chem 1999;274:38147-38154.</p> <p>Database Reference: INTERPRO; IPR002579;</p> <p>Comment: This domain has no known function. It is found associated</p> <p>Comment: with the peptide methionine sulfoxide reductase enzymatic domain PMSR. The domain has two conserved cysteine and histidines that could suggest and zinc binding site.</p> <p>Comment: The final cysteine is found to be replaced by the rare amino</p> <p>Comment: acid selenocysteine in some members of the family [1].</p> <p>Number of members: 26</p>
DUF26		Domain of unknown function DUF26	<p>Accession number: PF01657</p> <p>Definition: Domain of unknown function DUF26</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_980 (release 4.1)</p> <p>Gathering cutoffs: -8 -8</p> <p>Trusted cutoffs: 6.50 1.40</p> <p>Noise cutoffs: -17.50 -17.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Database Reference: INTERPRO; IPR002902;</p> <p>Database reference: PFAMB; PB005223;</p> <p>Comment: This domain has no known function. It is found in serine/threonine</p> <p>Comment: kinases, associated with the Eukaryotic protein kinase domain</p> <p>Comment: pkinase. In the 33kDa secretary protein Swiss:082551</p> <p>Comment: this domain is duplicated. The domain contains four conserved</p> <p>Comment: cysteines.</p> <p>Number of members: 25</p>
Dynein_light	PDOC00953	Dynein light chain type 1 signature	<p>Dynein is a multisubunit microtubule-dependent motor enzyme that acts as the force generating protein of eukaryotic cilia and flagella. The cytoplasmic isoform of dynein acts as a motor for the intracellular retrograde motility of vesicles and organelles along microtubules. Dynein is composed of a number of ATP-binding large subunits, intermediate size subunits and small subunits.</p> <p>Among the small subunits, there is a family [1,2] of highly conserved proteins which consist of:</p> <ul style="list-style-type: none"> - Chlamydomonas reinhardtii flagellar outer arm dynein 8 Kd and 11 Kd light chains. - Higher eukaryotes cytoplasmic dynein light chain 1. - Yeast cytoplasmic dynein light chain 1 (gene DYN2 or SLC1). - Caenorhabditis elegans hypothetical dynein light chains M18.2 and T26A5.9. <p>These proteins are have from 89 to 120 amino acids. As a signature pattern, we selected a highly conserved region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern H-x-l-x-G-[KR]-x-F-[GA]-S-x-V-[ST]-[HY]-E</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p>

886

			<p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / First entry.</p> <p>References [1] King S.M., Patel-King R.S. J. Biol. Chem. 270:11445-11452(1995).</p> <p>[2] Dick T., Ray K., Salz H.K., Chia W. Mol. Cell. Biol. 16:1966-1977(1996).</p>
eIF5_eIF2B		Domain found in IF2B/IF5	<p>Accession number: PF01873</p> <p>Definition: Domain found in IF2B/IF5</p> <p>Author: Enright A, Ouzounis C, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Enright A</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 233.00 233.00</p> <p>Noise cutoffs: -56.10 -56.10</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96060092</p> <p>Reference Title: Multidomain organization of eukaryotic guanine nucleotide</p> <p>Reference Title: exchange translation initiation factor eIF-2B subunits revealed by analysis of conserved sequence motifs.</p> <p>Reference Author: Koonin EV;</p> <p>Reference Location: Protein Sci 1995;4:1608-1617.</p> <p>Database Reference: INTERPRO; IPR002735;</p> <p>Comment: This family includes the N terminus of eIF-5</p> <p>Swiss:P55010, and</p> <p>Comment: the C terminus of eIF-2 beta Swiss:P20042. This region corresponds to the whole of the archaeobacterial eIF-2 beta</p> <p>Comment: homolog. The region contains a putative zinc binding C4 finger.</p> <p>Number of members: 20</p>
eIF6		eIF-6 family	<p>This family comprises members exhibiting sequence identity to the eukaryotic translation initiation factor 6. Some members of this family are implicated in protein biosynthesis as a translation initiation factor by binding to the 60s ribosomal subunit and preventing its association with the 40s ribosomal subunit to form the 80s initiation complex. Such activity can play a role in maximal polysome formation and plays an important role in determining free 60s ribosomal subunit content. Polypeptides in this family can optimize amino acid and nitrogen content in a desired cell or organism. References describing eif6 family members and their biological activities include, for example, the following: Adams et al., Science 87:2185-2195(2000); Wood et al., J. Biol. Chem. 274:11653-11659(1999); and Si et al., Mol. Cell. Biol. 19:1416-1426(1999).</p>
ER	PDOC00992	Enhancer of rudimentary signature	<p>The Drosophila protein 'enhancer of rudimentary' (gene (e(r)) is a small protein of 104 residues whose function is not yet clear. From an evolutionary point of view, it is highly conserved [1] and has been found to exist in probably all multicellular eukaryotic organisms. It has been proposed that this protein plays a role in the cell cycle.</p> <p>As as signaure pattern, we selected a conserved region in the central part of the protein.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern Y-D-I-[SA]-x-L-[FY]-x-F-[IV]-D-x(3)-D-[LIV]-S</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / First entry.</p> <p>References [1] Gelsthorpe M., Pulumati M., McCallum C., Dang-Vu K., Tsubota S.I. Gene 186:189-195(1997).</p>

ER_lumen_re cept	PDOC00732	ER lumen protein retaining receptor signatures	<p>Proteins that reside in the lumen of the endoplasmic reticulum (ER) contain a C-terminal tetrapeptide (generally K-D-E-L or H-D-E-L) that serves as a signal for their retrieval (retrograde transport) from subsequent compartments of the secretory pathway. The signal is recognized by a receptor molecule that is believed to cycle between the cis side of the Golgi apparatus and the ER [1]. This protein is known as the ER lumen protein retaining receptor or also as the 'KDEL receptor'. It has been characterized in a variety of species, including fungi (gene ERD2), plants, Plasmodium, Drosophila and mammals. In mammals two highly related forms of the receptor are known.</p> <p>Structurally, the receptor is a protein of about 220 residues that seems to contain seven transmembrane regions [2]. The N-terminal part (3 residues) is oriented toward the lumen while the C-terminal tail (about 12 residues) is cytoplasmic. There are three luminal and three cytoplasmic loops.</p> <p>We developed two signature patterns for these receptors. The first pattern corresponds to the C-terminal half of the first cytoplasmic loop as well as most of the second transmembrane domain. The second pattern is a perfectly conserved decapeptide that corresponds to the central part of the fifth transmembrane domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-[LIV]-S-x-[KR]-x-[QH]-x-L-[FY]-x-[LIV](2)-[FYW]-x(2)-R-Y Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern L-E-[SA]-V-A-I-[LM]-P-Q-[LI] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update December 1999 / Patterns and text revised.</p> <p>References [1] Pelham H.R.B. Curr. Opin. Cell Biol. 3:585-591(1991).</p> <p>[2] Townsend F.M., Wilson D.W., Pelham H.R.B. EMBO J. 12:2821-2829(1993).</p>
ETF_alpha	PDOC00583	Electron transfer flavoprotein alpha-subunit signature	<p>The electron transfer flavoprotein (ETF) [1,2] serves as a specific electron acceptor for various mitochondrial dehydrogenases. ETF transfers electrons to the main respiratory chain via ETF-ubiquinone oxidoreductase. ETF is an heterodimer that consist of an alpha and a beta subunit and which bind one molecule of FAD per dimer. A similar system also exists in some bacteria.</p> <p>The alpha subunit of ETF is a protein of about 32 Kd which is structurally related to the bacterial nitrogen fixation protein fixB which could play a role in a redox process and feed electrons to ferredoxin.</p> <p>Other related proteins are:</p> <ul style="list-style-type: none"> - Escherichia coli hypothetical protein ydiR. - Escherichia coli hypothetical protein ygcQ. <p>As a signature pattern for these proteins we selected a highly conserved region which is located in the C-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LI]-Y-[LIVM]-[AT]-x-G-[IV]-[SD]-G-x-[IV]-Q-H-x(2)-G-x(6)-[IV]-x-A-[IV]-N Sequences known to belong to this class detected by the pattern ALL, except for ygcQ.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1998 / Text revised.</p>

			<p>References</p> <p>[1] Finocchiaro G., Ikeda Y., Ito M., Tanaka K. Prog. Clin. Biol. Res. 321:637-652(1990).</p> <p>[2] Tsai M.H., Saier M.H. Jr. Res. Microbiol. 146:397-404(1995).</p>
Euk_porin	PDOC00483	Eukaryotic mitochondrial porin signature	<p>The major protein of the outer mitochondrial membrane of eukaryotes is a porin that forms a voltage-dependent anion-selective channel (VDAC) that behaves as a general diffusion pore for small hydrophilic molecules [1 to 4]. The channel adopts an open conformation at low or zero membrane potential and a closed conformation at potentials above 30-40 mV.</p> <p>This protein contains about 280 amino acids and its sequence is composed of between 12 to 16 beta-strands that span the mitochondrial outer membrane. Yeast contains two members of this family (genes POR1 and POR2); vertebrates have at least three members (genes VDAC1, VDAC2 and VDAC3) [5].</p> <p>As a signature pattern we selected a conserved region located at the C-terminal part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [YH]-x(2)-D-[SPCAD]-x-[STA]-x(3)-[TAG]-[KR]-[LIVMF]-[DNSTA]-[DNS]-x(4)-[GSTAN]-[LIVMA]-x-[LIVMY] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update July 1999 / Pattern and text revised.</p> <p>References</p> <p>[1] Benz R. Biochim. Biophys. Acta 1197:167-196(1994).</p> <p>[2] Manella C.A. Trends Biochem. Sci. 17:315-320(1992).</p> <p>[3] Dihanich M. Experientia 46:146-153(1990).</p> <p>[4] Forte M., Guy H.R., Mannella C.A. J. Bioenerg. Biomembr. 19:341-350(1987).</p> <p>[5] Sampson M.J., Lovell R.S., Davison D.B., Craigen W.J. Genomics 36:192-196(1996).</p>
F_bp_aldolase	PDOC00523	Fructose-bisphosphate aldolase class-II signatures	<p>Fructose-bisphosphate aldolase (EC 4.1.2.13) [1,2] is a glycolytic enzyme that catalyzes the reversible aldol cleavage or condensation of fructose-1,6-bisphosphate into dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate.</p> <p>There are two classes of fructose-bisphosphate aldolases with different catalytic mechanisms. Class-II aldolases [2], mainly found in prokaryotes and fungi, are homodimeric enzymes which require a divalent metal ion - generally zinc - for their activity.</p> <p>This family also includes the following proteins:</p> <ul style="list-style-type: none"> - Escherichia coli galactitol operon protein gatY which catalyzes the transformation of tagatose 1,6-bisphosphate into glycerone phosphate and D-glyceraldehyde 3-phosphate. - Escherichia coli N-acetyl galactosamine operon protein agaY which catalyzes the same reaction as that of gatY. <p>As signature patterns for this class of enzyme, we selected two conserved</p>

889

			<p>regions. The first pattern is located in the first half of the sequence and contains two histidine residues that have been shown [4] to be involved in binding a zinc ion. The second is located in the C-terminal section and contains clustered acidic residues and glycines.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [FYVMT]-x(1,3)-[LIVMH]-[APNT]-[LIVM]-x(1,2)-[LIVM]-H-x-D-H-[GACH] [The two H's are zinc ligands] Sequences known to belong to this class detected by the pattern ALL, except for <i>Mycoplasma pneumoniae</i> aldolase. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [LIVM]-E-x-E-[LIVM]-G-x(2)-[GM]-[GSTA]-x-E Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update December 1999 / Pattern and text revised.</p> <p>References [1] Perham R.N. Biochem. Soc. Trans. 18:185-187(1990).</p> <p>[2] Marsh J.J., Lebherz H.G. Trends Biochem. Sci. 17:110-113(1992).</p> <p>[3] von der Osten C.H., Barbas C.F. III, Wong C.-H., Sinskey A.J. Mol. Microbiol. 3:1625-1637(1989).</p> <p>[4] Berry A., Marshall K.E. FEBS Lett. 318:11-16(1993).</p>
FAA_hydrolase		Fumarylacetoacetate (FAA) hydrolase family	<p>Accession number: PF01557 Definition: Fumarylacetoacetate (FAA) hydrolase family Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_641 (release 4.0) Gathering cutoffs: 25 25 Trusted cutoffs: 42.10 42.10 Noise cutoffs: -93.10 -93.10 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1] Reference Medline: 97255958 Reference Title: Mutations in the fumarylacetoacetate hydrolase gene causing Reference Title: hereditary tyrosinemia type I: overview. Reference Author: St-Louis M, Tanguay RM; Reference Location: Hum Mutat 1997;9:291-299.</p> <p>Reference Number: [2] Reference Medline: 96125235 Reference Title: Molecular characterization of the 4-hydroxyphenylacetate catabolic pathway of <i>Escherichia coli</i> W: engineering a mobile aromatic degradative cluster. Reference Title: mobile aromatic degradative cluster. Reference Author: Prieto MA, Diaz E, Garcia JL; Reference Location: J Bacteriol 1996;178:111-120.</p> <p>Reference Number: [3] Reference Medline: 96016123 Reference Title: Fungal metabolic model for human type I hereditary tyrosinaemia. Reference Title: tyrosinaemia. Reference Author: Fernandez-Canon JM, Penalva MA; Reference Location: Proc Natl Acad Sci U S A 1995;92:9132-9136.</p> <p>Reference Number: [4] Reference Medline: 94039092 Reference Title: Purification, nucleotide sequence and some properties of a Reference Title: bifunctional isomerase/decarboxylase from the homoprotocatechuate degradative pathway of <i>Escherichia coli</i></p>

890

			<p>Reference Title: C.</p> <p>Reference Author: Roper DI, Cooper RA;</p> <p>Reference Location: Eur J Biochem 1993;217:575-580.</p> <p>Database reference: MIM; 276700;</p> <p>Database Reference: INTERPRO; IPR002529;</p> <p>Comment: This family consists of fumarylacetoacetate (FAA) hydrolase,</p> <p>Comment: or fumarylacetoacetate hydrolase (FAH) and it also includes</p> <p>Comment: HHDD isomerase/OPET decarboxylase from E. coli strain W.</p> <p>Comment: FAA is the last enzyme in the tyrosine catabolic pathway, it hydrolyses</p> <p>Comment: fumarylacetoacetate into fumarate and acetoacetate which then join the</p> <p>Comment: citric acid cycle [1]. Mutations in FAA cause type I tyrosinemia in humans</p> <p>Comment: this is an inherited disorder mainly affecting the liver leading to</p> <p>Comment: liver cirrhosis, hepatocellular carcinoma, renal tubular damages and</p> <p>Comment: neurologic crises amongst other symptoms [1]. The enzymatic defect causes</p> <p>Comment: the toxic accumulation of phenylalanine/tyrosine catabolites [3].</p> <p>Comment: The E. coli W enzyme HHDD isomerase/OPET decarboxylase contains two</p> <p>Comment: copies of this domain and functions in fourth and fifth steps of the</p> <p>Comment: homoprotocatechuate pathway;</p> <p>Comment: here it decarboxylates OPET to HHDD and isomerizes this to OHED.</p> <p>Comment: The final products of this pathway are pyruvic acid and succinic</p> <p>Comment: semialdehyde.</p> <p>Number of members: 33</p>
FAD_binding		FAD binding domain	<p>Accession number: PF00667</p> <p>Definition: FAD binding domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_180 (release 2.1)</p> <p>Gathering cutoffs: 16.8 16.8</p> <p>Trusted cutoffs: 24.60 16.80</p> <p>Noise cutoffs: 13.50 15.90</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95386502</p> <p>Reference Title: The flavin reductase activity of the flavoprotein component</p> <p>Reference Title: of sulfite reductase from Escherichia coli. A new model for</p> <p>Reference Title: the protein structure.</p> <p>Reference Author: Eschenbrenner M, Coves J, Fontecave M;</p> <p>Reference Location: J Biol Chem 1995;270:20550-20555.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 96049560</p> <p>Reference Title: NADPH-sulfite reductase flavoprotein from Escherichia coli:</p> <p>Reference Title: contribution to the flavin content and subunit interaction.</p> <p>Reference Author: Eschenbrenner M, Coves J, Fontecave M;</p> <p>Reference Location: FEBS Lett 1995;374:82-84.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 94360001</p> <p>Reference Title: Dissection of NADPH-cytochrome P450 oxidoreductase into</p> <p>Reference Title: distinct functional domains.</p> <p>Reference Author: Smith GC, Tew DG, Wolf CR;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1994;91:8710-8714.</p> <p>Reference Number: [4]</p> <p>Reference Medline: 97385116</p> <p>Reference Title: Three-dimensional structure of NADPH-cytochrome P450 reductase: prototype for FMN- and FAD-containing</p>

			<p>enzymes.</p> <p>Reference Author: Wang M, Roberts DL, Paschke R, Shea TM, Masters BS, Kim JJ;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1997;94:8411-8416.</p> <p>Database Reference: SCOP; 1amo; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR001709;</p> <p>Database Reference: PDB; 1amo A; 274; 493;</p> <p>Database Reference: PDB; 1amo B; 274; 493;</p> <p>Database Reference: PDB; 1quf ; 77; 120;</p> <p>Database reference: PFAMB; PB001390;</p> <p>Comment: This domain is found in sulfite reductase, NADPH cytochrome P450</p> <p>Comment: reductase and Nitric oxide synthase.</p> <p>Number of members: 87</p>
FAD_binding_3		FAD binding domain	<p>Accession number: PF01494</p> <p>Definition: FAD binding domain</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_549 (release 4.0)</p> <p>Gathering cutoffs: -7 -7</p> <p>Trusted cutoffs: -6.20 -6.20</p> <p>Noise cutoffs: -7.90 -7.90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 93028353</p> <p>Reference Title: Crystal structure of the reduced form of p-hydroxybenzoate</p> <p>Reference Title: hydroxylase refined at 2.3A resolution.</p> <p>Reference Author: Schreuder HA, van der Laan JM, Swarte MB, Kalk KH, Hol WG,</p> <p>Reference Author: Drenth J;</p> <p>Reference Location: Proteins 1992;14:178-190.</p> <p>Database Reference: SCOP; 2phh; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002938;</p> <p>Database Reference: PDB; 1pxa ; 5; 35;</p> <p>Database Reference: PDB; 1bf3 ; 5; 139;</p> <p>Database Reference: PDB; 1bgj ; 5; 139;</p> <p>Database Reference: PDB; 1bgn ; 5; 139;</p> <p>Database Reference: PDB; 1bkx ; 5; 139;</p> <p>Database Reference: PDB; 1cc4 A; 5; 139;</p> <p>Database Reference: PDB; 1cc6 A; 5; 139;</p> <p>Database Reference: PDB; 1cj2 A; 5; 139;</p> <p>Database Reference: PDB; 1pbb ; 5; 139;</p> <p>Database Reference: PDB; 1pbc ; 5; 139;</p> <p>Database Reference: PDB; 1pbd ; 5; 139;</p> <p>Database Reference: PDB; 1pbe ; 5; 139;</p> <p>Database Reference: PDB; 1pbf ; 5; 139;</p> <p>Database Reference: PDB; 1pdh ; 5; 139;</p> <p>Database Reference: PDB; 2phh ; 5; 139;</p> <p>Database Reference: PDB; 1cj3 A; 5; 139;</p> <p>Database Reference: PDB; 1cj4 A; 5; 139;</p> <p>Database Reference: PDB; 1phh ; 5; 139;</p> <p>Database Reference: PDB; 1d7l A; 5; 139;</p> <p>Database Reference: PDB; 1dob ; 5; 139;</p> <p>Database Reference: PDB; 1doc ; 5; 139;</p> <p>Database Reference: PDB; 1dod ; 5; 139;</p> <p>Database Reference: PDB; 1doe ; 5; 139;</p> <p>Database Reference: PDB; 1ius ; 5; 139;</p> <p>Database Reference: PDB; 1iut ; 5; 139;</p> <p>Database Reference: PDB; 1iuu ; 5; 139;</p> <p>Database Reference: PDB; 1iuv ; 5; 139;</p> <p>Database Reference: PDB; 1iuw ; 5; 139;</p> <p>Database Reference: PDB; 1iux ; 5; 139;</p> <p>Database Reference: PDB; 1foh A; 10; 151;</p> <p>Database Reference: PDB; 1foh D; 10; 151;</p> <p>Database Reference: PDB; 1foh B; 10; 151;</p> <p>Database Reference: PDB; 1foh C; 10; 151;</p> <p>Database reference: PFAMB; PB040546;</p> <p>Comment: This domain is involved in FAD binding in a number of enzymes.</p> <p>Number of members: 52</p>

892

FAD_binding _4	PDOC00674	Oxygen oxidoreductas es covalent FAD-binding site	<p>Some oxygen-dependent oxidoreductases are flavoproteins that contains a covalently bound FAD group which is attached to a histidine via an 8-alpha-(N3-histidyl)-riboflavin linkage. These proteins are:</p> <ul style="list-style-type: none"> - 6-hydroxy-D-nicotine oxidase (EC 1.5.3.6) (6-HDNO) [1], a bacterial enzyme that catalyzes the oxygen-dependent degradation of 6-hydroxynicotine into 6-hydroxypyrid-N-methylsamine - Plant reticuline oxidase (EC 1.5.3.9) [2] (berberine-bridge-forming enzyme), an enzyme that catalyzes the oxidation of (S)-reticuline into (S)-scoulerine in the pathway leading to benzophenanthridine alkaloids. - L-gulonolactone oxidase (EC 1.1.3.8) (l-gulono-gamma-lactone oxidase) [3], a mammalian enzyme which catalyzes the oxidation of L-gulono-1,4-lactone to L-xylo-hexulonolactone which spontaneously isomerizes to L-ascorbate. - D-arabinono-1,4-lactone oxidase (EC 1.1.3.24) (L-galactonolactone oxidase), a yeast enzyme involved in the biosynthesis of D-erythroascorbic acid [4]. - Mitomycin radical oxidase [5], a bacterial protein involved in mitomycin resistance and that probably oxidizes the reduced form of mitomycins. - Rhodococcus fascians fasciation locus protein fas5. <p>The region around the histidine that binds the FAD group is conserved in these enzymes and can be used as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-x(10)-[DE]-[LIVM]-x(3)-[LIVM]-x(9)-[LIVM]-x(3)-[GSA]-[GST]-G-H [H is the FAD binding site]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Text revised. EMBL/GenBank: U40390. References</p> <p>[1] Brandsch R., Hinkkanen A.E., Mauch L., Nagursky H., Decker K. Eur. J. Biochem. 167:315-320(1987).</p> <p>[2] Dittrich H., Kutchan T.M. Proc. Natl. Acad. Sci. U.S.A. 88:9969-9973(1991).</p> <p>[3] Koshizaka T., Nishikimi M., Ozawa T., Yagi K. J. Biol. Chem. 263:1619-1621(1988).</p> <p>[4] Huh W.-K., Kim S.-T., Kim J.-Y., Hwang S.-W., Kang S.-O.</p> <p>[5] August P.R., Flickinger M.C., Sherman D.H. J. Bacteriol. 176:4448-4454(1994).</p>
fer2	PDOC00175; PDOC00642	2Fe-2S ferredoxins, iron-sulfur binding region signature; Adrenodoxin family, iron- sulfur binding region signature	<p>Ferredoxins [1] are a group of iron-sulfur proteins which mediate electron transfer in a wide variety of metabolic reactions. Ferredoxins can be divided into several subgroups depending upon the physiological nature of the iron sulfur cluster(s) and according to sequence similarities. One of these subgroups are the 2Fe-2S ferredoxins, which are proteins or domains of around one hundred amino acid residues that bind a single 2Fe-2S iron-sulfur cluster. The proteins that are known [2] to belong to this family are listed below.</p> <ul style="list-style-type: none"> - Ferredoxin from photosynthetic organisms; namely plants and algae where it is located in the chloroplast or cyanelle; and cyanobacteria. - Ferredoxin from archaeobacteria of the Halobacterium genus. - Ferredoxin IV (gene pftA) and V (gene fdxD) from Rhodobacter capsulatus. - Ferredoxin in the toluene degradation operon (gene xylT) and naphthalene degradation operon (gene nahT) of Pseudomonas putida. - Hypothetical Escherichia coli protein yfaE. - The N-terminal domain of the bifunctional ferredoxin/ferredoxin reductase electron transfer component of the benzoate 1,2-dioxygenase complex (gene benC) from Acinetobacter calcoaceticus, the toluene 4-monooxygenase

		<p>complex (gene tmoF), the toluate 1,2-dioxygenase system (gene xylZ), and the xylene monooxygenase system (gene xylA) from <i>Pseudomonas</i>.</p> <ul style="list-style-type: none"> - The N-terminal domain of phenol hydroxylase protein p5 (gene dmpP) from <i>Pseudomonas Putida</i>. - The N-terminal domain of methane monooxygenase component C (gene mmoC) from <i>Methylococcus capsulatus</i>. - The C-terminal domain of the vanillate degradation pathway protein vanB in a <i>Pseudomonas</i> species. - The N-terminal domain of bacterial fumarate reductase iron-sulfur protein (gene frdB). - The N-terminal domain of CDP-6-deoxy-3,4-glucoseen reductase (gene ascD) from <i>Yersinia pseudotuberculosis</i>. - The central domain of eukaryotic succinate dehydrogenase (ubiquinone) iron-sulfur protein. - The N-terminal domain of eukaryotic xanthine dehydrogenase. - The N-terminal domain of eukaryotic aldehyde oxidase. <p>In the 2Fe-2S ferredoxins, four cysteine residues bind the iron-sulfur cluster. Three of these cysteines are clustered together in the same region of the protein. Our signature pattern spans that iron-sulfur binding region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-[C]-[C]-[GA]-[C]-C-[GAST]-[CPDEKRHFYW]-C [The three C's are 2Fe-2S ligands] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 15.</p> <p>Note in addition to the proteins listed above there are a number of other ferredoxin-like proteins that bind a 2Fe-2S cluster but which do not seem to be evolutionary related to this family. Among them are the ferredoxins from the adrenodoxin family (see <PDOC00642>) as well as the bacterial aromatic dioxygenase systems ferredoxin-like proteins such as bnzC, ndoA, and todB.</p> <p>Last update November 1997 / Text revised.</p> <p>References [1] Meyer J. Trends Ecol. Evol. 3:222-226(1988).</p> <p>[2] Harayama S., Polissi A., Rekik M. FEBS Lett. 285:85-88(1991).</p> <p>Ferredoxins [1] are a group of iron-sulfur proteins which mediate electron transfer in a wide variety of metabolic reactions. Ferredoxins can be divided into several subgroups depending upon the physiological nature of the iron sulfur cluster(s) and according to sequence similarities. One family of ferredoxins groups together the following proteins that all bind a single 2Fe-2S iron-sulfur cluster:</p> <ul style="list-style-type: none"> - Adrenodoxin (ADX) (adrenal ferredoxin), a vertebrate mitochondrial protein which transfers electrons from adrenodoxin reductase to cytochrome P450_{sc}, which is involved in cholesterol side chain cleavage. - Putidaredoxin (PTX), a <i>Pseudomonas putida</i> protein which transfers electrons from putidaredoxin reductase to cytochrome P450_{cam}, which is involved in the oxidation of camphor. - Terpredoxin [2], a <i>Pseudomonas</i> protein which transfers electrons from terpredoxin reductase to cytochrome P450_{terp}, which is involved in the oxidation of alpha-terpineol. - Rhodocoxin [3], a <i>Rhodococcus</i> protein which transfers electrons from rhodocoxin reductase to cytochrome CYP116 (thcB), which is involved in the degradation of thiocarbamate herbicides. - <i>Escherichia coli</i> ferredoxin (gene fdx) [4] whose exact function is not yet known. - <i>Rhodobacter capsulatus</i> ferredoxin VI [5], which may transfer electrons to a yet uncharacterized oxygenase. - <i>Caulobacter crescentus</i> ferredoxin (gene fdxB) [6].
--	--	---

894

			<p>In these proteins, four cysteine residues bind the iron-sulfur cluster. Three of these cysteines are clustered together in the same region of the protein. Our signature pattern spans that iron-sulfur binding region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-x(2)-[STAQ]-x-[STAMV]-C-[STA]-T-C-[HR] [The three C's are 2Fe-2S ligands] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 1. Last update November 1995 / Pattern and text revised. EMBL/Genbank: X51607. References [1] Meyer J. Trends Ecol. Evol. 3:222-226(1988).</p> <p>[2] Peterson J.A., Lu J.-Y., Geisselsoder J., Graham-Lorence S., Carmona C., Witney F., Lorence M.C. J. Biol. Chem. 267:14193-14203(1992).</p> <p>[3] Nagy I., Schoofs G., Compennolle F., Proost P., Vanderleyden J., De Mot R. J. Bacteriol. 177:676-687(1995).</p> <p>[4] Ta D.T., Vickery L.E. J. Biol. Chem. 267:11120-11125(1992).</p> <p>[5] Naud I., Vincon M., Garin J., Gaillard J., Forest E., Jouanneau Y. Eur. J. Biochem. 222:933-939(1994).</p> <p>[6] Amemiya K</p>
Ferric_reduct		Ferric reductase like transmembrane component	<p>Accession number: PF01794 Definition: Ferric reductase like transmembrane component Author: Bashton M, Bateman A Alignment method of seed: T_Coffee Source of seed members: Pfam-B_728 (release 4.2) Gathering cutoffs: -122 -122 Trusted cutoffs: -34.80 -34.80 Noise cutoffs: -210.30 -210.30 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 93309468 Reference Title: The fission yeast ferric reductase gene frp1+ is required for ferric iron uptake and encodes a protein that is homologous to the gp91-phox subunit of the human NADPH Reference Title: phagocyte oxidoreductase. Reference Author: Roman DG, Dancis A, Anderson GJ, Klausner RD; Reference Location: Mol Cell Biol 1993;13:4342-4350. Reference Number: [2] Reference Medline: 92294876 Reference Title: Cytochrome b558: the flavin-binding component of the phagocyte NADPH oxidase. Reference Author: Rotrosen D, Yeung CL, Leto TL, Malech HL, Kwong CH; Reference Location: Science 1992;256:1459-1462. Reference Number: [3] Reference Medline: 87258189 Reference Title: The glycoprotein encoded by the X-linked chronic granulomatous disease locus is a component of the neutrophil cytochrome b complex. Reference Title: neutrophil cytochrome b complex. Reference Author: Dinanier MC, Orkin SH, Brown R, Jesaitis AJ, Parkos CA; Reference Location: Nature 1987;327:717-720.</p>

895

			<p>Reference Number: [4] Reference Medline: 87258190 Reference Title: The X-linked chronic granulomatous disease gene codes for Reference Title: the beta- chain of cytochrome b-245. Reference Author: Teahan C, Rowe P, Parker P, Totty N, Segal AW; Reference Location: Nature 1987;327:720-721. Database Reference: INTERPRO; IPR002916; Comment: This family includes a common region in the transmembrane proteins Comment: mammalian cytochrome B-245 heavy chain (gp91-phox), ferric reductase Comment: transmembrane component in yeast and respiratory burst oxidase from Comment: mouse-ear cress. Comment: This may be a family of flavocytochromes capable of moving electrons Comment: across the plasma membrane [1]. Comment: The Frp1 protein Swiss:Q04800 from <i>S. pombe</i> is a ferric reductase Comment: component and is required for cell surface ferric reductase activity. Comment: mutants in <i>frp1</i> are deficient in ferric iron uptake [1]. Comment: Cytochrome B-245 heavy chain Swiss:P04839 is a FAD- dependent Comment: dehydrogenase it is also has electron transferase activity which reduces Comment: molecular oxygen to superoxide anion, a precursor in the production of Comment: microbicidal oxidants [2]. Comment: Mutations in the sequence of cytochrome B-245 heavy chain (gp91-phox) Comment: lead to the X-linked chronic granulomatous disease. The bacteriocidal Comment: ability of phagocytic cells is reduced and is characterised by the Comment: absence of a functional plasma membrane associated NADPH oxidase [3]. Comment: The chronic granulomatous disease gene codes for the beta chain of Comment: cytochrome B-245 and cytochrome B-245 is missing from patients with Comment: the disease [4]. Comment: The aligned region includes a potential FAD binding domain. Number of members: 34</p>
Flavi_NS5		Flavivirus RNA-directed RNA polymerase	<p>Accession number: PF00972 Definition: Flavivirus RNA-directed RNA polymerase Author: Finn RD, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_200 (release 3.0) Gathering cutoffs: 12 12 Trusted cutoffs: 16.00 16.00 Noise cutoffs: 8.50 8.50 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 95159427 Reference Title: Phylogeny of TYU, SRE, and CFA virus: different evolutionary rates in the genus Flavivirus. Reference Author: Marin MS, Zanotto PM, Gritsun TS, Gould EA; Reference Location: Virology 1995;206:1133-1139. Reference Number: [2] Reference Medline: 96182933 Reference Title: Recombinant dengue type 1 virus NS5 protein expressed in Reference Title: Escherichia coli exhibits RNA-dependent RNA polymerase Reference Title: activity. Reference Author: Tan BH, Fu J, Sugrue RJ, Yap EH, Chan YC, Tan YH; Reference Location: Virology 1996;216:317-325. Reference Number: [3] Reference Medline: 93224895</p>

			<p>Reference Title: Computer-assisted identification of a putative methyltransferase domain in NS5 protein of flaviviruses and</p> <p>Reference Title: lambda 2 protein of reovirus.</p> <p>Reference Author: Koonin EV;</p> <p>Reference Location: J Gen Virol 1993;74:733-740.</p> <p>Reference Number: [4]</p> <p>Reference Medline: 94094568</p> <p>Reference Title: Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences.</p> <p>Reference Title: Koonin EV, Dolja VV;</p> <p>Reference Author: Crit Rev Biochem Mol Biol 1993;28:375-430.</p> <p>Database Reference: INTERPRO; IPR000208;</p> <p>Comment: Flaviviruses produce a polyprotein from the ssRNA genome.</p> <p>Comment: This protein is also known as NS5.</p> <p>Comment: This RNA-directed RNA polymerase possesses a number of short</p> <p>Comment: regions and motifs homologous to other RNA-directed RNA</p> <p>Comment: polymerases [2].</p> <p>Number of members: 159</p>
Fork_head	PDOC00564	Fork head domain signatures and profile	<p>It has been shown [1] that some eukaryotic transcription factors contain a conserved domain of about 100 amino-acid residues, called the fork head domain (but also known as a "winged helix"), which is involved in DNA-binding [2]. Proteins known to contain this domain are listed below.</p> <ul style="list-style-type: none"> - Drosophila fork head protein (fkh). Fkh is probably a transcription factor that regulates the expression of genes involved in terminal development. - Drosophila protein crocodile (gene croc) [3], which is required for the establishment of head structures. - Drosophila proteins FD2, FD3, FD4, and FD5. - Drosophila proteins sloppy paired 1 and 2 (slp1 and slp2) involved in segmentation. - Bombyx mori silk gland factor-1 (SGF-1) which regulates transcription of the sericin-1 gene. - Mammalian transcriptional activators HNF-3-alpha, -beta, and -gamma. The HNF-3 proteins interact with the cis-acting regulatory regions of a number of liver genes. - Mammalian interleukin-enhancer binding factor (ILF). ILF binds to the purine-rich NFAT-like motifs in the HIV-1 LTR and the interleukin-2 promoter. ILF may be involved in both positive and negative regulation of important viral and cellular promoter elements. - Mammalian transcription factor BF-1 which plays an important role in the establishment of the regional subdivision of the developing brain and in the development of the telencephalon. - Human HTLF, a protein that binds to the purine-rich region in human T-cell leukemia virus long terminal repeat (HTLV-I LTR). - Mammalian transcription factors FREAC-1 (FKHL5, HFH-8), FREAC-2 (FKHL6), FREAC-3 (FKHL7, FKH-1), FREAC-4 (FKHL8), FREAC-5 (FKHL9, FKH-2, HFH-6), FREAC-6 (FKHL10, HFH-5), FREAC-7 (FKHL11), FREAC-8 (FKHL12, HFH-7), FKH-3, FKH-4, FKH-5, HFH-1 and HFH-4. - Human AFX1 which is involved in a chromosomal translocation that causes acute leukemia. - Human FKHR which is involved in a chromosomal translocation that causes rhabdomyosarcoma. - Xenopus XFKH1, a protein essential for normal axis formation. - Caenorhabditis elegans lin-31; involved in the regulation of vulval cell fates. - Yeast HCM1, a protein of unknown function. - Yeast FKH1. - Yeast FKH2. <p>The fork domain is highly conserved. We have developed two patterns for its detection. The first corresponds to the N-terminal section of the domain; the second is a heptapeptide located in the central section of the domain.</p>

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KR]-P-[PTQ]-[FYLVQH]-S-[FY]-x(2)-[LIVM]-x(3,4)-[AC]-[LIM]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for AFX1 and FKHR.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern W-[QKR]-[INS]-S-[LIV]-R-H</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References [1] Weigel D., Jaeckle H. Cell 63:455-456(1990).</p> <p>[2] Clark K.L., Halay E.D., Lai E., Burley S.K. Nature 364:412-420(1993).</p> <p>[3] Haecker U., Kaufmann E., Hartmann C., Juergens G., Knoechel W., Jaeckle H. EMBO J. 14:5306-5317(1995).</p>
FtsJ		FtsJ cell division protein	<p>Accession number: PF01728</p> <p>Definition: FtsJ cell division protein</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1791 (release 4.1)</p> <p>Gathering cutoffs: -38 -38</p> <p>Trusted cutoffs: -20.90 -20.90</p> <p>Noise cutoffs: -56.70 -56.70</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 93186701</p> <p>Reference Title: The Escherichia coli FtsH protein is a prokaryotic member of a protein family of putative ATPases involved in membrane functions, cell cycle control, and gene expression.</p> <p>Reference Title: expression.</p> <p>Reference Author: Tomoyasu T, Yuki T, Morimura S, Mori H, Yamanaka K, Niki H,</p> <p>Reference Author: Hiraga S, Ogura T;</p> <p>Reference Location: J Bacteriol 1993;175:1344-1351.</p> <p>Database Reference INTERPRO: IPR002877;</p> <p>Database reference: PFAMB; PB030182;</p> <p>Comment: This family consists of FtsJ from various bacterial and archaeal sources</p> <p>Comment: In E. coli FtsJ is not essential for growth but affects cell division [1].</p> <p>Number of members: 25</p>
FTSW ROD A_SPOVE	PDOC00352	Cell cycle proteins ftsW / rodA / spoVE signature	<p>A number of prokaryotic proteins involved in cell cycle processes have been found [1,2] to be structurally related, these proteins are:</p> <ul style="list-style-type: none"> - Escherichia coli and related bacteria cell division protein ftsW. This protein plays a role in the stabilization of the ftsZ ring during cell division. - Escherichia coli and related bacteria rod shape-determining protein rodA (or mrdB). It is required for the expression of the enzymatic activity of PBP2, which is thought to participate in the synthesis of peptidoglycan during the initiation of cell elongation. - Bacillus subtilis stage V sporulation protein E (spoVE). The exact function of spoVE in endospore formation is not known. - Bacillus subtilis hypothetical protein ylaO. - Bacillus subtilis hypothetical protein ywcF (ipa-42D). - Cyanophora paradoxa cyanelle ftsW homolog. This protein may be involved in the organelle division process. <p>All these proteins are hydrophobic integral membrane protein and contain about</p>

			<p>400 residues. We have selected the best conserved region, which is located in the C-terminal section, as a signature pattern for these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [NV]-x(5)-[GTR]-[LIVMA]-x-P-[PTLIVM]-x-G-[LIVM]-x(3)-[LIVMFW](2)-S-[YSA]-G-G-[STN]-[SA]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Ikeda M., Sato T., Wachi M., Jung H.K., Ishino F., Kobayashi Y., Matsuhashi M. J. Bacteriol. 171:6375-6378(1989).</p> <p>[2] Joris B., Dive G., Henriques A., Piggot P.J., Ghuysen J.-M. Mol. Microbiol. 4:513-517(1990).</p>
Furin-like		Furin-like cysteine rich region	<p>Members of this family include receptors that mediate transmembrane signalling. These receptors can bind to a number of factors including: amphiregulin, epidermal growth factor, gp30, heparin-binding egf, insulin, insulin-like growth factor I and II, neuregulins, transforming growth factor-alpha and, and vaccinia virus growth</p> <p>Signal transduction is mediated by catalytic activity of tyrosine kinase, such as ATP + A protein tyrosine = ADP + protein tyrosine phosphate. Typically, such signal transduction have been implicated in metabolic and developmental changes, including cell fate and differentiation. Examples include instruction of follicle cells to follow a dorsal pathway of development rather than the default ventral pathway. may also bind the spitz protein. References describing these family members and their biological activities:</p> <p>Abbot et al., J. Biol. Chem. 267:10759-10763(1992); Araki et al., J. Biol. Chem. 262:16186-16191(1987); Aroian et al., EMBO J. 13:360-366(1994); Aroian et al., Nature 348:693-699(1990); Barbetti et al., Diabetes 41:408-415(1992); Bargmann et al., Nature 319:226-230(1986); Cama et al., J. Biol. Chem. 268:8060-8069(1993); Cama et al., J. Clin. Endocrinol. Metab. 73:894-901(1991); Carrera et al., Hum. Mol. Genet. 2:1437-1441(1993); Clifford et al., Genetics 137:531-550(1994); Coccozza et al., Diabetes 41:521-526(1992); Cooke et al., Biochem. Biophys. Res. Commun. 177:1113-1120(1991); Coussens et al., Science 230:1132-1139(1985); Dickens et al., Biochem. Biophys. Res. Commun. 186:244-250(1992); Ebina et al., Cell 40:747-758(1985); Ebina et al., Proc. Natl. Acad. Sci. U.S.A. 84:704-708(1987); Ehsani et al., Genomics 15:426-429(1993); Elbein et al., Diabetes 42:429-434(1993); Elbein, Diabetes 38:737-743(1989); Fujita-Yamaguchi et al., Protein Seq. Data Anal. 1:3-6(1987); Gullick et al., EMBO J. 11:43-48(1992); Haruta et al., Diabetes 42:1837-1844(1993); Hubbard et al., EMBO J. 16:5572-5581(1997); Hubbard et al., Nature 372:746-754(1994); Iwanishi et al., Diabetologia 36:414-422(1993); Kadowaki et al., J. Clin. Invest. 86:254-264(1990); Kadowaki et al., Science 240:787-790(1988); Kim et al., Diabetologia 35:261-266(1992); Klinkhamer et al., EMBO J. 8:2503-2507(1989); Kusari et al., J. Biol. Chem. 266:5260-5267(1991); Lai et al., Neuron 6:691-704(1991); Lax et al., Mol. Cell. Biol. 8:1970-1978(1988); Lebrun et al., J. Biol. Chem. 268:11272-11277(1993); Lee et al., Oncogene 8:3403-3410(1993); Lesokhin et al., Dev. Biol. 205:129-144(1999); Livneh et al., Cell 40:599-607(1985). Longo et al., Proc. Natl. Acad. Sci. U.S.A. 90:60-64(1993); McKeon et al., Mol. Endocrinol. 4:647-656(1990); Moller et al., J. Biol. Chem. 265:14979-14985(1990); Moller et al., Mol. Endocrinol. 4:1183-1191(1990); Odawara et al., Science 245:66-68(1989); Raz et al., Genetics 129:191-201(1991). Sakai et al., J. Mol. Biol. 256:548-555(1996); Schaeffer et al., Biochem. Biophys. Res. Commun. 189:650-653(1992); Schejter et al., Cell 46:1091-1101(1986); Seino et al., Biochem. Biophys. Res. Commun. 159:312-316(1989); Seino et al., Diabetes 39:123-128(1990); Semba et al., Proc. Natl. Acad. Sci. U.S.A. 82:6497-6501(1985); Shier et al., J. Biol. Chem. 264:14605-14608(1989); Taira et al., Science 245:63-66(1989); Tewari et al., J. Biol. Chem. 264:16238-16245(1989); Ullrich et al., Nature 313:756-761(1985). Ullrich et al., EMBO J. 5:2503-2512(1986); van der Vorm et al., Diabetologia 36:172-174(1993); van der Vorm et al., J. Biol. Chem. 267:66-71(1992); Wadsworth et al., Nature 314:178-180(1985); White et al., Cell 54:641-</p>

			649(1988); Xu et al., J. Biol. Chem. 265:18673-18681(1990); Yamamoto et al., Nature 319:230-234(1986); and Yoshimasa et al., Science 240:784-787(1988).
Galactosyl_T		Galactosyltransferase	<p>Accession number: PF01762</p> <p>Definition: Galactosyltransferase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_885 (release 4.2)</p> <p>Gathering cutoffs: -46 -46</p> <p>Trusted cutoffs: -43.90 -43.90</p> <p>Noise cutoffs: -49.80 -49.80</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98079080</p> <p>Reference Title: Cloning of a human</p> <p>Reference Title: UDP-galactose:2-acetamido-2-deoxy-D-glucose 3beta-galactosyltransferase catalyzing the formation of type 1 chains.</p> <p>Reference Author: Kolbinger F, Streiff MB, Katopodis AG;</p> <p>Reference Location: J Biol Chem 1998;273:433-440.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98079027</p> <p>Reference Title: Genomic cloning and expression of three murine</p> <p>Reference Title: UDP-galactose: beta-N- acetylglucosamine</p> <p>Reference Title: beta1,3-galactosyltransferase genes.</p> <p>Reference Author: Hennes T, Dinter A, Kuhnert P, Mattu TS, Rudd PM, Berger</p> <p>Reference Author: EG;</p> <p>Reference Location: J Biol Chem 1998;273:58-65.</p> <p>Database Reference: INTERPRO; IPR002659;</p> <p>Database reference: PFAM; PB005938;</p> <p>Database reference: PFAM; PB012965;</p> <p>Comment: This family includes the galactosyltransferases</p> <p>Comment: UDP-galactose:2-acetamido-2-deoxy-D-glucose3beta-galactosyltransferase</p> <p>Comment: Swiss:O43825 [1] and UDP-Gal:beta-GlcNAc beta 1,3-galactosyltransferase</p> <p>Comment: Swiss:O54904 [2].</p> <p>Comment: Specific galactosyltransferases transfer galactose to GlcNAc terminal</p> <p>Comment: chains in the synthesis of the lacto-series oligosaccharides types 1</p> <p>Comment: and 2 [1].</p> <p>Number of members: 29</p>
G-alpha		G-protein alpha subunit	<p>Accession number: PF00503</p> <p>Definition: G-protein alpha subunit</p> <p>Author: Finn RD</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_11 (release 1.0)</p> <p>Gathering cutoffs: 13.8 13.8</p> <p>Trusted cutoffs: 13.80 13.80</p> <p>Noise cutoffs: 9.70 12.70</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 94353239</p> <p>Reference Title: Structures of active conformations of Gi alpha 1 and the mechanism of GTP hydrolysis.</p> <p>Reference Author: Coleman DE, Berghuis AM, Lee E, Linder ME, Gilman AG,</p> <p>Reference Author: Sprang SR;</p> <p>Reference Location: Science 1994;265:1405-1412.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97004345</p> <p>Reference Title: How G proteins work: a continuing story.</p> <p>Reference Author: Coleman DE, Sprang SR;</p> <p>Reference Location: Trends Biochem Sci 1996;21:41-44.</p> <p>Database Reference: PRINTS; PR00318;</p> <p>Database Reference: SCOP; 1gia; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR001019;</p> <p>Database Reference: PDB; 1gia ; 34; 343;</p> <p>Database Reference: PDB; 1gil ; 34; 343;</p>

900

			<p>Database Reference PDB; 1as0 ; 32; 344;</p> <p>Database Reference PDB; 1gfi ; 33; 345;</p> <p>Database Reference PDB; 1as2 ; 32; 346;</p> <p>Database Reference PDB; 1bh2 ; 32; 346;</p> <p>Database Reference PDB; 1cip A; 32; 347;</p> <p>Database Reference PDB; 1git ; 32; 348;</p> <p>Database Reference PDB; 1agr D; 11; 353;</p> <p>Database Reference PDB; 1gg2 A; 6; 348;</p> <p>Database Reference PDB; 1gp2 A; 6; 348;</p> <p>Database Reference PDB; 1bof ; 10; 353;</p> <p>Database Reference PDB; 1as3 ; 9; 353;</p> <p>Database Reference PDB; 1gdd ; 9; 353;</p> <p>Database Reference PDB; 1agr A; 6; 353;</p> <p>Database Reference PDB; 1tag ; 27; 340;</p> <p>Database Reference PDB; 1tad A; 27; 342;</p> <p>Database Reference PDB; 1tad B; 27; 342;</p> <p>Database Reference PDB; 1tnd B; 27; 342;</p> <p>Database Reference PDB; 1tnd C; 27; 342;</p> <p>Database Reference PDB; 1tad C; 27; 344;</p> <p>Database Reference PDB; 1tnd A; 27; 349;</p> <p>Database Reference PDB; 1cjk C; 39; 388;</p> <p>Database Reference PDB; 1cjt C; 39; 388;</p> <p>Database Reference PDB; 1cju C; 39; 388;</p> <p>Database Reference PDB; 1cju C; 39; 388;</p> <p>Database Reference PDB; 1azt A; 35; 391;</p> <p>Database Reference PDB; 1azt B; 35; 391;</p> <p>Database Reference PDB; 1azs C; 36; 393;</p> <p>Database reference: PFAMB; PB034080;</p> <p>Comment: G proteins couple receptors of extracellular signals to intracellular signaling pathways.</p> <p>Comment: The G protein alpha subunit binds guanyl nucleotide and is a weak GTPase.</p> <p>Number of members: 245</p>
GCV_H		Glycine cleavage H-protein	<p>Accession number: PF01597</p> <p>Definition: Glycine cleavage H-protein</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_988 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 27.90 27.90</p> <p>Noise cutoffs: -58.80 -58.80</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 94255425</p> <p>Reference Title: X-ray structure determination at 2.6-A resolution of a lipoate- containing protein: the H-protein of the glycine decarboxylase complex from pea leaves.</p> <p>Reference Author: Pares S, Cohen-Addad C, Sieker L, Neuburger M, Douce R;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1994;91:4850-4853.</p> <p>Database Reference: SCOP; 1hpc; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002930;</p> <p>Database Reference: PDB; 1hpc A; 2; 127;</p> <p>Database Reference: PDB; 1hpc B; 2; 127;</p> <p>Database Reference: PDB; 1hpc ; 2; 127;</p> <p>Comment: This is a family of glycine cleavage H-proteins, part of the glycine cleavage multienzyme complex (GCV) found in bacteria and the mitochondria of eukaryotes. GCV catalyses the catabolism of glycine in eukaryotes.</p> <p>Comment: A lipoyl group is attached to a completely conserved lysine residue.</p> <p>Comment: The H protein shuttles the methylamine group of glycine from the P protein to the T protein.</p> <p>Number of members: 40</p>
GCV_T		Glycine cleavage T-	<p>Accession number: PF01571</p> <p>Definition: Glycine cleavage T-protein (aminomethyl transferase)</p>

901

		protein (aminomethyl transferase)	<p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_933 (release 4.0)</p> <p>Gathering cutoffs: -146 -146</p> <p>Trusted cutoffs: -124.50 -124.50</p> <p>Noise cutoffs: -167.90 -167.90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97199363</p> <p>Reference Title: Cloning, and molecular characterization of the GCV1 gene</p> <p>Reference Title: encoding the glycine cleavage T-protein from <i>Saccharomyces cerevisiae</i>.</p> <p>Reference Author: McNeil JB, Zhang F, Taylor BV, Sinclair DA, Pearlman RE,</p> <p>Reference Author: Bognar AL;</p> <p>Reference Location: Gene 1997;186:13-20.</p> <p>Database Reference INTERPRO; IPR002536;</p> <p>Database reference: PFAMB; PB004229;</p> <p>Comment: This is a family of glycine cleavage T-proteins, part of the glycine</p> <p>Comment: cleavage multienzyme complex (GCV) found in bacteria and the mitochondria</p> <p>Comment: of eukaryotes. GCV catalyses the catabolism of glycine in eukaryotes.</p> <p>Comment: The T-protein is an aminomethyl transferase.</p> <p>Number of members: 27</p>
G-gamma	PDOC01002	G-protein gamma subunit profile	<p>Guanine nucleotide-binding proteins (G proteins) [1] act as intermediaries in the transduction of signals generated by transmembrane receptors. G proteins consist of three subunits (alpha, beta, and gamma). The alpha subunit binds to and hydrolyzes GTP; the functions of the beta and gamma subunits are less clear but they seem to be required for the replacement of GDP by GTP as well as for membrane anchoring and receptor recognition.</p> <p>The gamma subunits are small proteins (from 70 to 110 residues) that are bound to the membrane via a isoprenyl group (either a farnesyl or a geranyl-geranyl) covalently linked to their C-terminus. In mammals there are at least 12 different isoforms of gamma subunits.</p> <p>The <i>Caenorhabditis elegans</i> protein egl-10, which is a regulator of G-protein signalling, contains a G-protein gamma-like domain.</p> <p>We have developed a profile that spans the complete length of the gamma subunit.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Sequences known to belong to this class detected by the profile ALL, except for yeast and squid G-protein gamma.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Expert(s) to contact by email Pennington S.R. srpenn@liverpool.ac.uk</p> <p>Last update November 1997 / First entry.</p> <p>References [1] Pennington S.R. Protein Prof. 2:16-315(1995).</p>
glutaredoxin	PDOC00173	Glutaredoxin	<p>Glutaredoxin [1,2,3], also known as thioltransferase, is a small protein of approximately one hundred amino-acid residues. It functions as an electron carrier in the glutathione-dependent synthesis of deoxyribonucleotides by the enzyme ribonucleotide reductase. Like thioredoxin, which functions in a similar way, glutaredoxin possesses an active center disulfide bond. It exists in either a reduced or an oxidized form where the two cysteine residues are linked in an intramolecular disulfide bond.</p> <p>Glutaredoxin has been sequenced in a variety of species. On the basis of</p>

902

			<p>extensive sequence similarity, it has been proposed [4] that vaccinia protein O2L is most probably a glutaredoxin. Finally, it must be noted that phage T4 thioredoxin seems also to be evolutionary related.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVD]-[FYSA]-x(4)-C-[PV]-[FYWH]-C-x(2)-[TAV]-x(2,3)-[LIV] [The two C's form the redox-active bond] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note in position 5 of the pattern, all glutaredoxin sequences have Pro while T4 thioredoxin has Val.</p> <p>Last update December 1999 / Pattern and text revised.</p> <p>References [1] Gleason F.K., Holmgren A. FEMS Microbiol. Rev. 54:271-298(1988).</p> <p>[2] Holmgren A. Biochem. Soc. Trans. 16:95-96(1988).</p> <p>[3] Holmgren A. J. Biol. Chem. 264:13963-13966(1989).</p> <p>[4] Johnson G.P., Goebel S.J., Perkus M.E., Davis S.W., Winslow J.P., Paoletti E. Virology 181:378-381(1991).</p>
Glyco_hydro_1	PDOC00495	Glycosyl hydrolases family 1 signatures	<p>It has been shown [1 to 4] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:</p> <ul style="list-style-type: none"> - Beta-glucosidases (EC 3.2.1.21) from various bacteria such as Agrobacterium strain ATCC 21400, Bacillus polymyxa, and Caldocellum saccharolyticum. - Two plants (clover) beta-glucosidases (EC 3.2.1.21). - Two different beta-galactosidases (EC 3.2.1.23) from the archaeobacteria Sulfolobus solfataricus (genes bgaS and lacS). - 6-phospho-beta-galactosidases (EC 3.2.1.85) from various bacteria such as Lactobacillus casei, Lactococcus lactis, and Staphylococcus aureus. - 6-phospho-beta-glucosidases (EC 3.2.1.86) from Escherichia coli (genes bglB and ascB) and from Erwinia chrysanthemi (gene arbB). - Plants myrosinases (EC 3.2.3.1) (sinigrinase) (thioglucosidase). - Mammalian lactase-phlorizin hydrolase (LPH) (EC 3.2.1.108 / EC 3.2.1.62). <p>LPH, an integral membrane glycoprotein, is the enzyme that splits lactose in the small intestine. LPH is a large protein of about 1900 residues which contains four tandem repeats of a domain of about 450 residues which is evolutionary related to the above glycosyl hydrolases.</p> <p>One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [5], in the beta-glucosidase from Agrobacterium, to be directly involved in glycosidic bond cleavage by acting as a nucleophile. We have used this region as a signature pattern. As a second signature pattern we selected a conserved region, found in the N-terminal extremity of these enzymes, this region also contains a glutamic acid residue.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMFSTC]-[LIVFYS]-[LIV]-[LIVMST]-E-N-G-[LIVMFAR]-[CSAGN] [E is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 12.</p> <p>Note this pattern will pick up the last two domains of LPH; the first two domains, which are removed from the LPH precursor by proteolytic processing, have lost the active site glutamate and may therefore be inactive [4].</p> <p>Consensus pattern F-x-[FYWM]-[GSTA]-x-[GSTA]-x-[GSTA](2)-[FYNH]-[NQ]-x-</p>

			<p>E-x- [GSTA] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this pattern will pick up the last three domains of LPH. Expert(s) to contact by email Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update November 1995 / Patterns and text revised.</p> <p>References [1] Henrissat B. Biochem. J. 280:309-316(1991).</p> <p>[2] Henrissat B. Protein Seq. Data Anal. 4:61-62(1991).</p> <p>[3] Gonzalez-Candelas L., Ramon D., Polaina J. Gene 95:31-38(1990).</p> <p>[4] El Hassouni M., Henrissat B., Chippaux M., Barras F. J. Bacteriol. 174:765-777(1992).</p> <p>[5] Withers S.G., Warren R.A.J., Street I.P., Rupitz K., Kempton J.B., Aebersold R. J. Am. Chem. Soc. 112:5887-5889(1990).</p>
Glyco_hydro_19	PDOC00620	Chitinases family 19 signatures	<p>Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1]. Chitinases of family 19 (also known as classes IA or I and IB or II) are enzymes from plants that function in the defense against fungal and insect pathogens by destroying their chitin-containing cell wall. Class IA/I and IB/II enzymes differ in the presence (IA/I) or absence (IB/II) of a N-terminal chitin-binding domain (see the relevant entry <PDOC00025>). The catalytic domain of these enzymes consist of about 220 to 230 amino acid residues.</p> <p>As signature patterns we selected two highly conserved regions, the first one is located in the N-terminal section and contains one of the six cysteines which are conserved in most, if not all, of these chitinases and which is probably involved in a disulfide bond.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-x(4,5)-F-Y-[ST]-x(3)-[FY]-[LIVMF]-x-A-x(3)-[YF]-x(2)-F-[GSA] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [LIVM]-[GSA]-F-x-[STAG](2)-[LIVMFY]-W-[FY]-W-[LIVM] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Expert(s) to contact by email Neuhaus J.-M. jean-marc.neuhaus@bota.unine.ch</p> <p>Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update November 1997 / Text revised.</p> <p>References [1] Flach J., Pilet P.-E., Jolles P. Experientia 48:701-716(1992).</p> <p>[2]</p>

			<p>Henrissat B. Biochem. J. 280:309-316(1991).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?glycosid.txt</p>
Glyco_hydro_3_C	PDOC00621	Glycosyl hydrolases family 3 active site	<p>It has been shown [1,2] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:</p> <ul style="list-style-type: none"> - Beta glucosidases (EC 3.2.1.21) from the fungi <i>Aspergillus wentii</i> (A-3), <i>Hansenula anomala</i>, <i>Kluyveromyces fragilis</i>, <i>Saccharomycopsis fibuligera</i>, (BGL1 and BGL2), <i>Schizophyllum commune</i> and <i>Trichoderma reesei</i> (BGL1). - Beta glucosidases from the bacteria <i>Agrobacterium tumefaciens</i> (Cbg1), <i>Butyrivibrio fibrilvolvens</i> (bglA), <i>Clostridium thermocellum</i> (bglB), <i>Escherichia coli</i> (bglX), <i>Erwinia chrysanthemi</i> (bgxA) and <i>Ruminococcus albus</i>. - <i>Alteromonas</i> strain O-7 beta-hexosaminidase A (EC 3.2.1.52). - <i>Bacillus subtilis</i> hypothetical protein yzbA. - <i>Escherichia coli</i> hypothetical protein ycfO and HI0959, the corresponding <i>Haemophilus influenzae</i> protein. <p>One of the conserved regions in these enzymes is centered on a conserved aspartic acid residue which has been shown [3], in <i>Aspergillus wentii</i> beta-glucosidase A3, to be implicated in the catalytic mechanism. We have used this region as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM](2)-[KR]-x-[EQK]-x(4)-G-[LIVMFT]-[LIVT]-[LIVMF]-[ST]-D-x(2)-[SGADNI] [D is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Expert(s) to contact by email Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Henrissat B. Biochem. J. 280:309-316(1991).</p> <p>[2] Castle L.A., Smith K.D., Morris R.O. J. Bacteriol. 174:1478-1486(1992).</p> <p>[3] Bause E., Legler G. Biochim. Biophys. Acta 626:459-465(1980).</p>
Glyco_hydro_45	PDOC00877	Glycosyl hydrolases family 45 active site	<p>The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family K or as the glycosyl hydrolases family 45 [3,E1]. The enzymes which are currently known to belong to this family are listed below.</p> <ul style="list-style-type: none"> - Endoglucanase 5 from <i>Humicola insolens</i>. - Endoglucanase 5 from <i>Trichoderma reesei</i> (egl5). - Endoglucanase K from <i>Fusarium oxysporum</i>. - Endoglucanase B from <i>Pseudomonas fluorescens</i> (celB). - Endoglucanase 1 from <i>Ustilago maydis</i> (egl1). <p>The best conserved regions in these enzymes is located in the N-terminal section. It contains an aspartic acid residue which has been shown [4] to act as a nucleophile in the catalytic mechanism. We use this region as a signature pattern.</p>

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [STA]-T-R-Y-[FYW]-D-x(5)-[CA] [The D is an active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Expert(s) to contact by email Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).</p> <p>[2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).</p> <p>[3] Henrissat B., Bairoch A. Biochem. J. 293:781-788(1993).</p> <p>[4] Davies G.J., Dodson G.G., Hubbard R.E., Tolley S.P., Dauter Z., Wilson K.S., Hjort C., Mikkelsen J.M., Rasmussen G., Schuelein M. Nature 365:362-364(1993).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?glycosid.txt</p>
Glyco_hydro_47		Glycosyl hydrolase family 47	<p>Members of this family are alpha-mannosidases that catalyse the hydrolysis of the terminal 1,2-linked alpha-D-mannose residues in the oligo-mannose oligosaccharide Man(9)(GlcNAc)(2). These enzymes are capable of taking part in the glycosylation pathway and glycoprotein processing.</p>
GTP_cyclohydrol	PDOC00672	GTP cyclohydrolase I signatures	<p>GTP cyclohydrolase I (EC 3.5.4.16) catalyzes the biosynthesis of formic acid and dihydroneopterin triphosphate from GTP. This reaction is the first step in the biosynthesis of tetrahydrofolate in prokaryotes, of tetrahydrobiopterin in vertebrates, and of pteridine-containing pigments in insects.</p> <p>GTP cyclohydrolase I is a protein of from 190 to 250 amino acid residues. The comparison of the sequence of the enzyme from bacterial and eukaryotic sources shows that the structure of this enzyme has been extremely well conserved throughout evolution [1].</p> <p>As signature patterns we selected two conserved regions. The first contains a perfectly conserved tetrapeptide which is part of the GTP-binding pocket [2], the second region also contains conserved residues involved in GTP-binding.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [DEN]-[LIVM](2)-x(2)-[KRNQ]-[DEN]-[LIVM]-x(3)-[ST]-x-C-E-H-H</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [SA]-x-[RK]-x-Q-[LIVM]-Q-E-[RN]-[LI]-[TSN]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1999 / Patterns and text revised.</p> <p>References [1] Maier J., Witter K., Guetlich M., Ziegler I., Werner T., Ninnemann H. Biochem. Biophys. Res. Commun. 212:705-711(1995).</p> <p>[2] Nar H., Huber R., Meining W., Schmid C., Weinkauff S., Bacher A.</p>

906

			Structure 3:459-466(1995).
HCV_capsid		Hepatitis C virus capsid protein	<p>Family members include nucleocapsid proteins of the HCV. This virus family comprises a nucleocapsid covered by a lipoprotein envelope. The envelope consists of two proteins: protein M and glycoprotein E. The nucleocapsid is a complex of protein c and mRNA. Uses for these polypeptides include: immunological epitopes for vaccines; or as mRNA chaperone proteins to aid in processing or to prevent degradation.</p> <p>References describing examples of these capsid polypeptides include: Chen et al., Virology 188:102-113(1992); and Okamoto et al., J. Gen. Virol. 72:2697-2704(1991)</p>
HD		HD domain	<p>Accession number: PF01966 Definition: HD domain Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: -1 -1 Trusted cutoffs: -0.50 -0.50 Noise cutoffs: -2.50 -2.50 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 99085258 Reference Title: The HD domain defines a new superfamily of metal-dependent phosphohydrolases. Reference Author: Aravind L, Koonin EV; Reference Location: Trends Biochem Sci 1998;23:469-472. Database Reference: INTERPRO; IPR002819; Database reference: PFAMB; PB005654; Database reference: PFAMB; PB006725; Database reference: PFAMB; PB009617; Database reference: PFAMB; PB012663; Database reference: PFAMB; PB035384; Database reference: PFAMB; PB040597; Comment: HD domains are metal dependent phosphohydrolases. Number of members: 63</p>
HDV_ag		Hepatitis delta virus delta antigen	<p>Accession number: PF01517 Definition: Hepatitis delta virus delta antigen Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_808 (release 4.0) Gathering cutoffs: -8 -8 Trusted cutoffs: 23.30 23.30 Noise cutoffs: -40.50 -40.50 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 94065676 Reference Title: Characterization of RNA-binding domains of hepatitis delta antigen. Reference Author: Poisson F, Roingeard P, Baillou A, Dubois F, Bonelli F, Calogero RA, Goudeau A; Reference Location: J Gen Virol 1993;74:2473-2478. Reference Number: [2] Reference Medline: 98362586 Reference Title: Structural basis of the oligomerization of hepatitis delta antigen. Reference Author: Zuccola HJ, Rozzelle JE, Lemon SM, Erickson BW, Hogle JM; Reference Location: Structure 1998;6:821-830. Database Reference: SCOP; 1a92; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002506; Database Reference: PDB; 1a92 A; 12; 23; Database Reference: PDB; 1a92 B; 12; 23; Database Reference: PDB; 1a92 C; 12; 23; Database Reference: PDB; 1a92 D; 12; 60; Database Reference: PDB; 1a92 A; 47; 60; Database Reference: PDB; 1a92 B; 47; 60; Database Reference: PDB; 1a92 C; 47; 60; Comment: The hepatitis delta virus (HDV) encodes a single protein,</p>

			<p>the</p> <p>Comment: hepatitis delta antigen (HDAg). The central region of this protein</p> <p>Comment: has been shown to bind RNA [1]. Several interactions are also</p> <p>Comment: mediated by a coiled-coil region at the N terminus of the protein [2].</p> <p>Number of members: 145</p>
hemolysinCa bind	PDOC00293	Hemolysin- type calcium- binding region signature	<p>Gram-negative bacteria produce a number of proteins which are secreted into the growth medium by a mechanism that does not require a cleaved N-terminal signal sequence. These proteins, while having different functions, seem [1] to share two properties: they bind calcium and they contain a variable number of tandem repeats consisting of a nine amino acid motif rich in glycine, aspartic acid and asparagine. It has been shown [2] that such a domain is involved in the binding of calcium ions in a parallel beta roll structure. The proteins which are currently known to belong to this category are:</p> <ul style="list-style-type: none"> - Hemolysins from various species of bacteria. Bacterial hemolysins are exotoxins that attack blood cell membranes and cause cell rupture. The hemolysins which are known to contain such a domain are those from: <i>E. coli</i> (gene hlyA), <i>A. pleuropneumoniae</i> (gene appA), <i>A. actinomycetemcomitans</i> and <i>P. haemolytica</i> (leukotoxin) (gene lktA). - Cyclolysin from <i>Bordetella pertussis</i> (gene cyaA). A multifunctional protein which is both an adenylate cyclase and a hemolysin. - Extracellular zinc proteases: serralyisin (EC 3.4.24.40) from <i>Serratia</i>, prtB and prtC from <i>Erwinia chrysanthemi</i> and aprA from <i>Pseudomonas aeruginosa</i>. - Nodulation protein nodO from <i>Rhizobium leguminosarum</i>. <p>We derived a signature pattern from conserved positions in the sequence of the calcium-binding domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern D-x-[LI]-x(4)-G-x-D-x-[LI]-x-G-G-x(3)-D</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this pattern is found once in nodO and the extracellular proteases but up to 5 times in some hemolysin/cyclolysins.</p> <p>Last update October 1993 / Text revised.</p> <p>References [1] Economou A., Hamilton W.D.O., Johnston A.W.B., Downie J.A. EMBO J. 9:349-354(1990).</p> <p>[2] Baumann U., Wu S., Flaherty K.M., McKay D.B. EMBO J. 12:3357-3364(1993).</p>
Herpes_alk_e xo		Herpesvirus alkaline exonuclease	<p>Accession number: PF01771</p> <p>Definition: Herpesvirus alkaline exonuclease</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_822 (release 4.2)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 318.00 318.00</p> <p>Noise cutoffs: -277.60 -277.60</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 85107093</p> <p>Reference Title: Studies on the herpes simplex virus alkaline nuclease:</p> <p>Reference Title: detection of type-common and type-specific epitopes on the</p> <p>Reference Title: enzyme.</p> <p>Reference Author: Banks LM, Halliburton IW, Purifoy DJ, Killington RA, Powell</p>

			<p>Reference Author: KL; Reference Location: J Gen Virol 1985;66:1-14. Database Reference: INTERPRO; IPR001616; Comment: This family includes various alkaline exonucleases from Comment: members of the herpesviridae. Alkaline exonuclease Comment: appears to have an important role in the replication of Comment: herpes simplex virus [1]. Number of members: 23</p>
Herpes_gl		Alphaherpesvirus glycoprotein I	<p>Accession number: PF01688 Definition: Alphaherpesvirus glycoprotein I Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1222 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 157.20 157.20 Noise cutoffs: -126.70 -126.70 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96357074 Reference Title: Biosynthesis of glycoproteins E and I of feline Reference Title: herpesvirus: gE-gl interaction is required for Reference Title: intracellular transport. Reference Author: Mijnes JD, van der Horst LM, van Anken E, Horzinek MC, Reference Author: Rottier PJ, de Groot RJ; Reference Location: J Virol 1996;70:5466-5475. Reference Number: [2] Reference Medline: 94267406 Reference Title: Identification of the feline herpesvirus type 1 (FHV-1) Reference Title: genes encoding glycoproteins G, D, I and E: expression Reference Title: of Reference Title: FHV-1 glycoprotein D in vaccinia and raccoon poxviruses. Reference Author: Spatz SJ, Rota PA, Maes RK; Reference Location: J Gen Virol 1994;75:1235-1244. Reference Number: [3] Reference Medline: 94267879 Reference Title: Unusual phosphorylation sequence in the gplV (gl) Reference Title: component Reference Title: of the varicella-zoster virus gpl-gplV glycoprotein complex Reference Title: (VZV gE-gl complex). Reference Author: Yao Z, Grose C; Reference Location: J Virol 1994;68:4204-4211. Database Reference: INTERPRO; IPR002874; Comment: This family consists of glycoprotein I form various Comment: members of the Comment: alphaherpesvirinae these include herpesvirus, varicella- Comment: zoster virus Comment: and pseudorabies virus. Glycoprotein I (gl) is important Comment: during natural Comment: infection, mutants lacking gl produce smaller lesions at the Comment: site of Comment: infection and show reduced neuronal spread [1]. gl forms Comment: a heterodimeric Comment: complex with gE; this complex displays Fc receptor activity Comment: (binds to Comment: the Fc region of immunoglobulin) [1]. Glycoproteins are Comment: also important Comment: in the production of virus-neutralizing antibodies and cell Comment: mediated Comment: immunity [2]. The alphaherpesvirinae have a dsDNA Comment: genome and have no Comment: RNA stage during viral replication. Number of members: 22</p>
Herpes_glyco p_D		Herpesvirus glycoprotein M	<p>Accession number: PF01528 Definition: Herpesvirus glycoprotein M Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_929 (release 4.0) Gathering cutoffs: 25 25 Trusted cutoffs: 197.30 197.30 Noise cutoffs: -229.70 -229.70</p>

909

			<p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96357105</p> <p>Reference Title: Identification and characterization of pseudorabies virus glycoprotein gM as a nonessential virion component.</p> <p>Reference Author: Dijkstra JM, Visser N, Mettenleiter TC, Klupp BG;</p> <p>Reference Location: J Virol 1996;70:5684-5688.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 95381611</p> <p>Reference Title: Identification and molecular characterization of the murine cytomegalovirus homolog of the human cytomegalovirus UL100</p> <p>Reference Title: gene.</p> <p>Reference Author: Li W, Eidman K, Gehrz RC, Kari B;</p> <p>Reference Location: Virus Res 1995;36:163-175.</p> <p>Database Reference: INTERPRO; IPR000785;</p> <p>Comment: The herpesvirus glycoprotein M (gM) is an integral membrane protein</p> <p>Comment: predicted to contain 8 transmembrane segments [2].</p> <p>Glycoprotein M is</p> <p>Comment: not essential for viral replication [1].</p> <p>Number of members: 24</p>
HesB-like	PDOC00887	Hypothetical hesB/yadR/yf hF family signature	<p>The following uncharacterized proteins have been shown [1] to share regions of similarities:</p> <ul style="list-style-type: none"> - Anabaena and related cyanobacteria protein hesB which may be required for nitrogen fixation. - Escherichia coli hypothetical protein yadR and HI1723, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein ydiC. - Escherichia coli hypothetical protein yfH and HI0376, the corresponding Haemophilus influenzae protein. - Mycobacterium tuberculosis hypothetical protein Rv2204c. - Synechocystis strain PCC 6803 hypothetical protein slr1417. - Synechocystis strain PCC 6803 hypothetical protein slr1565. - A hypothetical protein in the nifU 5' region of many nitrogen fixing bacteria. - Porphyra purpurea chloroplast hypothetical protein in apcF-rps4 intergenic region. - Yeast hypothetical protein YLL027W. - Yeast hypothetical protein YPR067W. <p>These are small proteins (106 to 135 amino-acid residues in bacteria, about 200 residues in fungi) that contain a number of conserved regions. The most noteworthy of these regions is located in the C-terminal extremity, it contains two conserved cysteines. We have used this region as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern F-x-[LIVMFY]-x-N-[PG]-[NSKQ]-x(4)-C-x-C-[GS]-x-S-F</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update</p> <p>December 1999 / Pattern and text revised.</p> <p>References</p> <p>[1]</p> <p>Bairoch A., Rudd K.E.</p> <p>Unpublished observations (1995).</p>
HisG	PDOC1020	ATP phosphoribosyltransferase signature	<p>ATP phosphoribosyltransferase (EC 2.4.2.17) is the enzyme that catalyzes the first step in the biosynthesis of histidine in bacteria, fungi and plants. It is a protein of about 23 to 32 Kd. As a signature pattern we selected a region located in the C-terminal part of this enzyme.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern E-x(5)-G-x-[SAG]-x(2)-[IV]-x-D-[LV]-x(2)-[ST]-G-x-T-[LM]</p>

910

			<p>Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update July 1998 / First entry.</p>
histone	<p>PDOC00045 PDOC00046 PDOC00287 PDOC00308</p>	<p>Histone H2A signature; Histone H4 signature; Histone H3 signatures; Histone H2B signature</p>	<p>Histone H2A is one of the four histones, along with H2B, H3 and H4, which forms the eukaryotic nucleosome core. Using alignments of histone H2A sequences [1,2,E1] we selected, as a signature pattern, a conserved region in the N-terminal part of H2A. This region is conserved both in classical S-phase regulated H2A's and in variant histone H2A's which are synthesized throughout the cell cycle.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [AC]-G-L-x-F-P-V Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 2. Last update November 1995 / Pattern and text revised. References [1] Wells D.E., Brown D. Nucleic Acids Res. 19:2173-2188(1991).</p> <p>[2] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).</p> <p>[E1] http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES/index.html</p> <p>Histone H4 is one of the four histones, along with H2A, H2B and H3, which forms the eukaryotic nucleosome core. Along with H3, it plays a central role in nucleosome formation. The sequence of histone H4 has remained almost invariant in more than 2 billion years of evolution [1,E1]. The region we use as a signature pattern is a pentapeptide found in positions 14 to 18 of all H4 sequences. It contains a lysine residue which is often acetylated [2] and a histidine residue which is implicated in DNA-binding [3].</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-A-K-R-H Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 1. Last update November 1995 / Text revised. References [1] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).</p> <p>[2] Doenecke D., Gallwitz D. Mol. Cell. Biochem. 44:113-128(1982).</p> <p>[3] Ebralidse K.K., Grachev S.A., Mirzabekov A.D. Nature 331:365-367(1988).</p> <p>[E1] http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES/index.html</p> <p>Histone H3 is one of the four histones, along with H2A, H2B and H4, which forms the eukaryotic nucleosome core. It is a highly conserved protein of 135 amino acid residues [1,2,E1].</p> <p>The following proteins have been found to contain a C-terminal H3-like domain:</p>

911

		<p>- Mammalian centromeric protein CENP-A [3]. Could act as a core histone necessary for the assembly of centromeres.</p> <p>- Yeast chromatin-associated protein CSE4 [4].</p> <p>- Caenorhabditis elegans chromosome III encodes two highly related proteins (F54C8.2 and F58A4.3) whose C-terminal section is evolutionary related to the last 100 residues of H3. The function of these proteins is not yet known.</p> <p>We developed two signature patterns, The first one corresponds to a perfectly conserved heptapeptide in the N-terminal part of H3. The second one is derived from a conserved region in the central section of H3.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern K-A-P-R-K-Q-L Sequences known to belong to this class detected by the pattern ALL, except for the H3-like proteins and some protozoan H3. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern P-F-x-[RA]-L-[VA]-[KRQ]-[DEG]-[IV] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References [1] Wells D.E., Brown D. Nucleic Acids Res. 19:2173-2188(1991).</p> <p>[2] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).</p> <p>[3] Sullivan K.F., Hechenberger M., Masri K. J. Cell Biol. 127:581-592(1994).</p> <p>[4] Stoler S., Keith K.C., Curnick K.E., Fitzgerald-Hayes M. Genes Dev. 9:573-586(1995).</p> <p>[E1] http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES/index.html</p> <p>Histone H2B is one of the four histones, along with H2A, H3 and H4, which forms the eukaryotic nucleosome core. Using alignments of histone H2B sequences [1,2,E1], we selected a conserved region in the C-terminal part of H2B.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KR]-E-[LIVM]-[EQ]-T-x(2)-[KR]-x-[LIVM](2)-x-[PAG]-[DE]-L-x-[KR]-H-A-[LIVM]-[STA]-E-G Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1995 / Pattern and text revised.</p> <p>References [1] Wells D.E., Brown D. Nucleic Acids Res. 19:2173-2188(1991).</p> <p>[2] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).</p> <p>[E1] http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES/index.html</p>
--	--	--

912

HMA	PDOC00804	Heavy-metal-associated domain	<p>A conserved domain of about 30 amino acid residues has been found [1] in a number of proteins that transport or detoxify heavy metals. This domain contains two conserved cysteines that could be involved in the binding of these metals. The domain has been termed Heavy-Metal-Associated (HMA). It has been found in:</p> <ul style="list-style-type: none"> - A variety of cation transport ATPases (E1-E2 ATPases) (see <PDOC00139>). The human copper ATPases ATP7A and ATP7B which are respectively involved in Menke's and Wilson's diseases. ATP7A and ATP7B both contain 6 tandem copies of the HMA domain. The copper ATPases CCC2 from budding yeast, copA from <i>Enterococcus faecalis</i> and synA from <i>Synechococcus</i> contain one copy of the HMA domain. The cadmium ATPases cadA from <i>Bacillus firmus</i> and from plasmid p1258 from <i>Staphylococcus aureus</i> also contain a single HMA domain, while a chromosomal <i>Staphylococcus aureus</i> cadA contains two copies. Other, less characterized ATPases that contain the HMA domain are: fixI from <i>Rhizobium meliloti</i>, pacS from <i>Synechococcus</i> strain PCC 7942), <i>Mycobacterium leprae</i> ctpA and ctpB and <i>Escherichia coli</i> hypothetical protein yhhO. In all these ATPases the HMA domain(s) are located in the N-terminal section. - Mercuric reductase (EC 1.16.1.1) (gene merA) which is generally encoded by plasmids carried by mercury-resistant Gram-negative bacteria. Mercuric reductase is a class-1 pyridine nucleotide-disulphide oxidoreductase (see <PDOC00073>). There is generally one HMA domain (with the exception of a chromosomal merA from <i>Bacillus</i> strain RC607 which has two) in the N-terminal part of merA. - Mercuric transport protein periplasmic component (gene merP), also encoded by plasmids carried by mercury-resistant Gram-negative bacteria. It seems to be a mercury scavenger that specifically binds to one Hg(2+) ion and which passes it to the mercuric reductase via the merT protein. The N-terminal half of merP is a HMA domain. - <i>Helicobacter pylori</i> copper-binding protein copP. - Yeast protein ATX1 [2], which could act in the transport and/or partitioning of copper. <p>The consensus pattern for HMA spans the complete domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVNS]-x(2)-[LIVMFA]-x-C-x-[STAGCDNH]-C-x(3)-[LIVFG]-x(3)-[LIV]-x(9,11)-[IVA]-x-[LVFYs] [The two C's probably bind metals] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 6. Last update December 1999 / Pattern and text revised. References [1] Bull P.C., Cox D.W. Trends Genet. 10:246-252(1994). [2] Lin S.-J., Culotta V.L. Proc. Natl. Acad. Sci. U.S.A. 92:3784-3788(1995).</p>
HMG-CoA_red	PDOC00064	Hydroxymethylglutaryl-coenzyme A reductase signatures and profile	<p>Hydroxymethylglutaryl-coenzyme A reductase (EC 1.1.1.34) (HMG-CoA reductase) [1,2] catalyzes the NADP-dependent synthesis of mevalonate from 3-hydroxy-3-methylglutaryl-CoA. In vertebrates, HMG-CoA reductase is the rate-limiting enzyme in cholesterol biosynthesis. In plants, mevalonate is the precursor of all isoprenoid compounds.</p> <p>HMG-CoA reductase is a membrane bound enzyme. Structurally, it consists of 3</p>

913

			<p>domains. An N-terminal region that contains a variable number of transmembrane segments (7 in mammals, insects and fungi; 2 in plants), a linker region and a C-terminal catalytic domain of approximately 400 amino-acid residues.</p> <p>In archeobacteria [3] HMG-CoA reductase, which is involved in the biosynthesis of the isoprenoids side chains of lipids, seems to be cytoplasmic and lack the N-terminal hydrophobic domain.</p> <p>Some bacteria, such as <i>Pseudomonas mevalonii</i>, can use mevalonate as the sole carbon source. These bacteria use an NAD-dependent HMG-CoA reductase (EC 1.1.1.88) to deacetylate mevalonate into 3-hydroxy-3-methylglutaryl-CoA [3]. The <i>Pseudomonas</i> enzyme is structurally related to the catalytic domain of NADP-dependent HMG-CoA reductases.</p> <p>We selected three conserved regions as signature patterns for HMG-CoA reductases. The first is located in the center of the catalytic domain, the second is a glycine-rich region located in the C-terminal section of the same catalytic domain and the third is also located in the C-terminal section and contains an histidine residue that seems [4] to be implicated in the catalytic mechanism as a general base.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [RKH]-x(6)-D-x-M-G-x-N-x-[LIVMA] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 4.</p> <p>Consensus pattern [LIVM]-G-x-[LIVM]-G-G-[AG]-T Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 5.</p> <p>Consensus pattern A-[LIVM]-x-[STAN]-x(2)-[LI]-x-[KRNQ]-[GSA]-H-[LM]-x-[FYLH] [H is an active site residue] Sequences known to belong to this class detected by the pattern ALL, except for archaeobacterial HMG-CoA reductases. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.</p> <p>Last update November 1997 / Patterns and text revised; profile added.</p> <p>References [1] Caelles C., Ferrer A., Balcells L., Hegardt F.G., Boronat A. Plant Mol. Biol. 13:627-638(1989).</p> <p>[2] Basson M.E., Thorsness M., Finer-Moore J., Stroud R.M., Rine J. Mol. Cell. Biol. 8:3797-3808(1988).</p> <p>[3] Lam W.L., Doolittle W.F. J. Biol. Chem. 267:5829-5834(1992).</p> <p>[4] Beach M.J., Rodwell V.W. J. Bacteriol. 171:2994-3001(1989).</p> <p>[5] Darnay B.G., Wang Y., Rodwell V.W. J. Biol. Chem. 267:15064-15070(1992).</p>
HMGL-like	PDOC00813 PDOC00643	Hydroxymethylglutaryl-coenzyme A	3-hydroxy-3-methylglutaryl-coenzyme A lyase (HMG-CoA lyase or HL) (EC 4.1.3.4) catalyzes the transformation of HMG-CoA into acetyl-CoA and acetoacetate. In

914

		lyase active site; Alpha-isopropylmalate and homocitrate synthases signatures	<p>vertebrates it is a mitochondrial enzyme which is involved in ketogenesis and in leucine catabolism [1]. In some bacteria, such as <i>Pseudomonas mevalonii</i>, it is involved in mevalonate catabolism (gene <i>mvaB</i>). A cysteine has been shown [2], in <i>mvaB</i>, to be required for the activity of the enzyme. The region around this residue is perfectly conserved and is used as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern S-V-A-G-L-G-G-C-P-Y [C is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update November 1995 / First entry. References [1] Mitchell G.A., Robert M.-F., Hruz P.W., Wang S., Fontaine G., Behnke C.E., Mende-Mueller L.M., Schappert K., Lee C., Gibson K.M., Miziorko H.M. J. Biol. Chem. 268:4376-4381(1993). [2] Hruz P.W., Narasimhan C., Miziorko H.M. Biochemistry 31:6842-6847(1992).</p> <p>The following enzymes have been shown [1] to be functionally as well as evolutionary related:</p> <ul style="list-style-type: none"> - Alpha-isopropylmalate synthase (EC 4.1.3.12) which catalyzes the first step in the biosynthesis of leucine, the condensation of acetyl-CoA and alpha-ketoisovalerate to form 2-isopropylmalate synthase. - Homocitrate synthase (EC 4.1.3.21) (gene <i>nifV</i>) which is involved in the biosynthesis of the iron-molybdenum cofactor of nitrogenase and catalyzes the condensation of acetyl-CoA and alpha-ketoglutarate into homocitrate. - Soybean late nodulin 56. - <i>Methanococcus jannaschii</i> hypothetical proteins MJ0503, MJ1195 and MJ1392. <p>We have selected two conserved regions as signature patterns for these enzymes. The first region is located in the N-terminal section while the second region is located in the central section and contains two conserved histidine residues which could be implicated in the catalytic mechanism.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern L-R-[DE]-G-x-Q-x(10)-K Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [LIVMFV]-x(2)-H-x-H-[DN]-D-x-G-x-[GAS]-x-[GASLI] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update November 1997 / Patterns and text revised. References [1] Wang S.-Z., Dean D.R., Chen J.-S., Johnson J.L. J. Bacteriol. 173:3041-3046(1991).</p>
hormone5	PDOC00237	Neurohypophyseal hormones signature	<p>Oxytocin (or ocytocin) and vasopressin [1] are small (nine amino acid residues), structurally and functionally related neurohypophyseal peptide hormones. Oxytocin causes contraction of the smooth muscle of the uterus and of the mammary gland while vasopressin has a direct antidiuretic action on the kidney and also causes vasoconstriction of the peripheral vessels. Like the majority of active peptides, both hormones are synthesized as larger protein precursors that are enzymatically converted to their mature forms. Peptides belonging to this family are also found in birds, fish, reptiles and amphibians (mesotocin, isotocin, valitocin, glutitocin, aspartocin, vasotocin, seritocin, asvatocin, phasvatocin), in worms (annetocin), octopi</p>

915

			<p>(cephalotocin), locust (locupressin or neuropeptide F1/F2) and in molluscs (conopressins G and S) [2].</p> <p>The pattern developed to detect this category of peptides spans their entire sequence and includes four invariant amino acid residues.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-[LIFY](2)-x-N-[CS]-P-x-G [The two C's are linked by a disulfide bond].</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1995 / Pattern and text revised.</p> <p>References [1] Acher R., Chauvet J. Biochimie 70:1197-1207(1988).</p> <p>[2] Chauvet J., Michel G., Ouedraogo Y., Chou J., Chait B.T., Acher R. Int. J. Pept. Protein Res. 45:482-487(1995).</p>
HPPK	PDOC00631	7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase signature	<p>All organisms require reduced folate cofactors for the synthesis of a variety of metabolites. Most microorganisms must synthesize folate de novo because they lack the active transport system of higher vertebrate cells which allows these organisms to use dietary folates. Enzymes involved in folate biosynthesis are therefore targets for a variety of antimicrobial agents such as trimethoprim or sulfonamides.</p> <p>7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (EC 2.7.6.3) (HPPK) catalyzes the attachment of pyrophosphate to 6-hydroxymethyl-7,8-dihydropterin to form 6-hydroxymethyl-7,8-dihydropteridine pyrophosphate. This is the first step in a three-step pathway leading to 7,8-dihydrofolate.</p> <p>Bacterial HPPK (gene folK or sulD) [1] is a protein of 160 to 270 amino acids. In the lower eukaryote <i>Pneumocystis carinii</i>, HPPK is the central domain of a multifunctional folate synthesis enzyme (gene fas) [2].</p> <p>As a signature for HPPK, we selected a conserved region located in the central section of these enzymes.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KRHD]-x-[GA]-[PSAE]-R-x(2)-D-[LIV]-D-[LIVM](2)</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1999 / Pattern and text revised.</p> <p>References [1] Talarico T.L., Ray P.H., Dev I.K., Merrill B.M., Dallas W.S. J. Bacteriol. 174:5971-5977(1992).</p> <p>[2] Volpes F., Dyer M., Scaife J.G., Darby G., Stammers D.K., Delves C.J. Gene 112:213-218(1992).</p>
Hydrolase		haloacid dehalogenase-like hydrolase	<p>Accession number: PF00702</p> <p>Definition: haloacid dehalogenase-like hydrolase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_566 (release 2.1)</p> <p>Gathering cutoffs: 7 7</p> <p>Trusted cutoffs: 7.10 7.10</p> <p>Noise cutoffs: 2.90 2.90</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p>

			<p>Reference Medline: 96355356</p> <p>Reference Title: Crystal structure of L-2-haloacid dehalogenase from <i>Pseudomonas</i> sp. YL. An alpha/beta hydrolase structure that</p> <p>Reference Title: is different from the alpha/beta hydrolase fold.</p> <p>Reference Author: Hisano T, Hata Y, Fujii T, Liu JQ, Kurihara T, Esaki N, Soda K;</p> <p>Reference Location: J Biol Chem 1996;271:20322-20330.</p> <p>Database Reference: SCOP; 1jud; sf; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR001454;</p> <p>Database Reference: PDB; 1jud ; 4; 197;</p> <p>Database Reference: PDB; 1zrm ; 4; 197;</p> <p>Database Reference: PDB; 1zrn ; 4; 197;</p> <p>Database Reference: PDB; 1aq6 A; 2; 193;</p> <p>Database Reference: PDB; 1aq6 B; 2; 193;</p> <p>Database Reference: PDB; 1qq5 A; 2; 193;</p> <p>Database Reference: PDB; 1qq5 B; 2; 193;</p> <p>Database Reference: PDB; 1qq6 A; 2; 193;</p> <p>Database Reference: PDB; 1qq6 B; 2; 193;</p> <p>Database Reference: PDB; 1qq7 A; 2; 193;</p> <p>Database Reference: PDB; 1qq7 B; 2; 193;</p> <p>Database Reference: PDB; 1cqz A; 4; 19;</p> <p>Database Reference: PDB; 1cr6 A; 4; 19;</p> <p>Database Reference: PDB; 1cqz B; 4; 206;</p> <p>Database Reference: PDB; 1cr6 B; 4; 206;</p> <p>Database Reference: PDB; 1cqz A; 48; 206;</p> <p>Database Reference: PDB; 1cr6 A; 48; 206;</p> <p>Database reference: PFAMB; PB000701;</p> <p>Database reference: PFAMB; PB001048;</p> <p>Database reference: PFAMB; PB019234;</p> <p>Database reference: PFAMB; PB032787;</p> <p>Database reference: PFAMB; PB040985;</p> <p>Database reference: PFAMB; PB041061;</p> <p>Database reference: PFAMB; PB041182;</p> <p>Database reference: PFAMB; PB041477;</p> <p>Database reference: PFAMB; PB041535;</p> <p>Database reference: PFAMB; PB041628;</p> <p>Database reference: PFAMB; PB041677;</p> <p>Comment: This family are structurally different from the alpha/beta hydrolase family (abhydrolase).</p> <p>Comment: This family includes L-2-haloacid dehalogenase, epoxide hydrolases and phosphatases.</p> <p>Comment: The structure of the family consists of two domains. One is an inserted four helix bundle, which is the least well conserved region of the alignment, between residues 16</p> <p>Comment: and</p> <p>Comment: 96 of Swiss:P24069. The rest of the fold is composed of the</p> <p>Comment: core alpha/beta domain.</p> <p>Number of members: 134</p>
HypB_UreG		HypB/UreG nucleotide-binding domain	<p>Accession number: PF01495</p> <p>Definition: HypB/UreG nucleotide-binding domain</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_428 (release 4.0)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 197.70 197.70</p> <p>Noise cutoffs: -40.00 -40.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97285753</p> <p>Reference Title: The HypB protein from <i>Bradyrhizobium japonicum</i> can store</p> <p>Reference Title: nickel and is required for the nickel-dependent transcriptional regulation of hydrogenase.</p> <p>Reference Author: Olson JW, Fu C, Maier RJ;</p> <p>Reference Location: Mol Microbiol 1997;24:119-128.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97352660</p> <p>Reference Title: Characterization of UreG, identification of a</p> <p>Reference Title: UreD-UreF-UreG complex, and evidencesuggesting that a</p>

			<p>Reference Number: [4] Reference Medline: 96270582 Reference Title: The binding site of karyopherin alpha for karyopherin beta overlaps with a nuclear localization sequence. Reference Author: Moroianu J, Blobel G, Radu A; Reference Location: Proc Natl Acad Sci U S A 1996;93:6572-6576. Reference Number: [5] Reference Medline: 96203101 Reference Title: A 41 amino acid motif in importin-alpha confers binding to importin- beta and hence transit into the nucleus. Reference Author: Gorlich D, Henklein P, Laskey RA, Hartmann E; Reference Location: EMBO J 1996;15:1810-1817. Database Reference: SCOP; 1bk5; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002652; Database Reference: PDB; 1ejl I; 72; 99; Database Reference: PDB; 1ejj I; 72; 99; Database Reference: PDB; 1ial A; 44; 99; Database Reference: PDB; 1qgr B; 28; 51; Database Reference: PDB; 1qgk B; 11; 54; Database Reference: PDB; 1ee5 A; 90; 110; Database Reference: PDB; 1bk5 A; 89; 110; Database Reference: PDB; 1bk5 B; 89; 110; Database Reference: PDB; 1bk6 A; 89; 110; Database Reference: PDB; 1bk6 B; 89; 110; Database Reference: PDB; 1ee4 A; 87; 110; Database Reference: PDB; 1ee4 B; 87; 110; Comment: This family consists of the importin alpha (karyopherin alpha), Comment: importin beta (karyopherin beta) binding domain. The domain mediates Comment: formation of the importin alpha beta complex; required for classical Comment: NLS import of proteins into the nucleus, through the nuclear pore Comment: complex and across the nuclear envelope. Comment: Also in the alignment is the NLS of importin alpha which overlaps Comment: with the IBB domain [4]. Number of members: 38</p>
IF-2B		Initiation factor 2 subunit family	<p>Accession number: PF01008 Definition: Initiation factor 2 subunit family Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1302 (release 3.0) Gathering cutoffs: -135 -135 Trusted cutoffs: -82.40 -82.40 Noise cutoffs: -157.30 -157.30 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98188271 Reference Title: Archaeal translation initiation revisited: the initiation factor 2 and eukaryotic initiation factor 2B Reference Title: alpha-beta-delta subunit families. Reference Author: Kyrpides NC, Woese CR; Reference Location: Proc Natl Acad Sci U S A 1998;95:3726-3730. Database Reference: INTERPRO; IPR000649; Comment: This family includes initiation factor 2B alpha, beta and delta Comment: subunits from eukaryotes, initiation factor 2B subunits 1 and 2 Comment: from archaeobacteria and some proteins of unknown function from Comment: prokaryotes. Initiation factor 2 binds to Met-tRNA, GTP and the Comment: small ribosomal subunit. Number of members: 33</p>
IF3	PDOC00723	Initiation factor 3 signature	<p>Initiation factor 3 (IF-3) (gene infC) [1] is one of the three factors required for the initiation of protein biosynthesis in bacteria. IF-3 is thought to function as a fidelity factor during the assembly of the ternary initiation complex which consist of the 30S ribosomal subunit, the initiator tRNA and the messenger RNA. IF-3 binds to the 30S ribosomal subunit; it is a</p>

			<p>basic protein of 141 to 212 residues.</p> <p>The chloroplast initiation factor IF-3(chl) is a protein that enhances the poly(A,U,G)-dependent binding of the initiator tRNA to chloroplast ribosomal 30s subunits. In its mature form it is a protein of about 400 residues whose central section is evolutionary related to the sequence of bacterial IF-3 [2].</p> <p>As a signature pattern we selected a highly conserved region located in the central section of bacterial IF-3 and of IF-3(chl).</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KR]-[LIVM](2)-[DN]-[FY]-[GSN]-[KR]-[LIVMFYS]-x-[FY]-[DEQTH]-x(2)-[KRQ]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1999 / Pattern and text revised.</p> <p>References [1] Liveris D., Schwartz J.J., Geertman R., Schwartz I. FEMS Microbiol. Lett. 112:211-216(1993).</p> <p>[2] Lin Q., Ma L., Burkhart W., Spremulli L.L. J. Biol. Chem. 269:9436-9444(1994).</p>
IF4E	PDOC00641	Eukaryotic initiation factor 4E signature	<p>Eukaryotic translation initiation factor 4E (eIF-4E) [1] is a protein that binds to the cap structure of eukaryotic cellular mRNAs. eIF-4E recognizes and binds the 7-methylguanosine-containing (m7Gppp) cap during an early step in the initiation of protein synthesis and facilitates ribosome binding to a mRNA by inducing the unwinding of its secondary structures.</p> <p>eIF-4E is a conserved protein of about 25 Kd. Site directed mutagenesis experiments have shown [2] that a tryptophan in the central part of the sequence of human eIF-4E seems to be implicated in cap-binding. The signature pattern for eIF-4E includes this tryptophan.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [DE]-[IFY]-x(2)-F-[KR]-x(2)-[LIVM]-x-P-x-W-E-[DVA]-x(5)-G-G-[KR]-W [The first W seems to be involved in cap-binding]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1999 / Pattern and text revised.</p> <p>References [1] Thach R.E. Cell 68:177-180(1992).</p> <p>[2] Ueda H., Iyo H., Doi M., Inoue M., Ishida T., Morioka H., Tanaka T., Nishikawa S., Uesugi S. FEBS Lett. 280:207-210(1991).</p>
IF5_eIF4_eIF2		eIF4-gamma/eIF5/eIF2-epsilon	<p>Accession number: PF02020</p> <p>Definition: eIF4-gamma/eIF5/eIF2-epsilon</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 26.10 26.10</p> <p>Noise cutoffs: -21.50 -21.50</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96060092</p> <p>Reference Title: Multidomain organization of eukaryotic guanine</p>

920

			<p>nucleotide</p> <p>Reference Title: exchange translation initiation factor eIF-2B subunits revealed by analysis of conserved sequence motifs.</p> <p>Reference Author: Koonin EV;</p> <p>Reference Location: Protein Sci 1995;4:1608-1617.</p> <p>Comment: This domain of unknown function is found at the C-terminus</p> <p>Comment: of several transcription initiation factors [1].</p> <p>Number of members: 31</p>
ig	PDOC00262	Immunoglobulins and major histocompatibility complex proteins signature	<p>The basic structure of immunoglobulin (Ig) [1] molecules is a tetramer of two light chains and two heavy chains linked by disulfide bonds. There are two types of light chains: kappa and lambda, each composed of a constant domain (CL) and a variable domain (VL). There are five types of heavy chains: alpha, delta, epsilon, gamma and mu, all consisting of a variable domain (VH) and three (in alpha, delta and gamma) or four (in epsilon and mu) constant domains (CH1 to CH4).</p> <p>The major histocompatibility complex (MHC) molecules are made of two chains.</p> <p>In class I [2] the alpha chain is composed of three extracellular domains, a transmembrane region and a cytoplasmic tail. The beta chain (beta-2-microglobulin) is composed of a single extracellular domain. In class II [3], both the alpha and the beta chains are composed of two extracellular domains, a transmembrane region and a cytoplasmic tail.</p> <p>It is known [4,5] that the Ig constant chain domains and a single extracellular domain in each type of MHC chains are related. These homologous domains are approximately one hundred amino acids long and include a conserved intradomain disulfide bond. We developed a small pattern around the C-terminal cysteine involved in this disulfide bond which can be used to detect these category of Ig related proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [FY]-x-C-x-[VA]-x-H-Sequences known to belong to this class detected by the pattern: Ig heavy chains type Alpha C region : All, in CH2 and CH3. Ig heavy chains type Delta C region : All, in CH3. Ig heavy chains type Epsilon C region: All, in CH1, CH3 and CH4. Ig heavy chains type Gamma C region : All, in CH3 and also CH1 in some cases Ig heavy chains type Mu C region : All, in CH2, CH3 and CH4. Ig light chains type Kappa C region : In all CL except rabbit and Xenopus. Ig light chains type Lambda C region : In all CL except rabbit. MHC class I alpha chains : All, in alpha-3 domains, including in the cytomegalovirus MHC-1 homologous protein [6]. Beta-2-microglobulin : All. MHC class II alpha chains: All, in alpha-2 domains. MHC class II beta chains: All, in beta-2 domains.</p> <p>Other sequence(s) detected in SWISS-PROT 71.</p> <p>Last update</p> <p>May 1991 / Text revised.</p> <p>References</p> <p>[1]</p> <p>Gough N.</p> <p>Trends Biochem. Sci. 6:203-205(1981).</p> <p>[2]</p> <p>Klein J., Figueroa F.</p> <p>Immunol. Today 7:41-44(1986).</p> <p>[3]</p> <p>Figueroa F., Klein J.</p> <p>Immunol. Today 7:78-81(1986).</p> <p>[4]</p> <p>Orr H.T., Lancet D., Robb R.J., Lopez de Castro J.A., Strominger J.L.</p> <p>Nature 282:266-270(1979).</p> <p>[5]</p> <p>Cushley W., Owen M.J.</p> <p>Immunol. Today 4:88-92(1983).</p> <p>[6]</p>

921

			Beck S., Barrel B.G. Nature 331:269-272(1988).
IMPDH_C	PDOC00391	IMP dehydrogenase / GMP reductase signature	<p>IMP dehydrogenase (EC 1.1.1.205) (IMPDH) catalyzes the rate-limiting reaction of de novo GTP biosynthesis, the NAD-dependent reduction of IMP into XMP [1]. Inhibition of IMP dehydrogenase activity results in the cessation of DNA synthesis. As IMP dehydrogenase is associated with cell proliferation, it is a possible target for cancer chemotherapy. Mammalian and bacterial IMPDHs are tetramers of identical chains. There are two IMP dehydrogenase isozymes in humans [2].</p> <p>GMP reductase (EC 1.6.6.8) catalyzes the irreversible and NADPH-dependent reductive deamination of GMP into IMP [3]. It converts nucleobase, nucleoside and nucleotide derivatives of G to A nucleotides, and maintains intracellular balance of A and G nucleotides.</p> <p>IMP dehydrogenase and GMP reductase share many regions of sequence similarity. One of these regions is centered on a cysteine residue thought [3] to be involved in binding IMP. We have used this region as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-[RK]-[LIVM]-G-[LIVM]-G-x-G-S-[LIVM]-C-x-T [C is the putative IMP-binding residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update May 1991 / First entry. References [1] Collart F.R., Huberman E. J. Biol. Chem. 263:15769-15772(1988). [2] Natsumeda Y., Ohno S., Kawasaki H., Konno Y., Weber G., Suzuki K. J. Biol. Chem. 265:5292-5295(1990). [3] Andrews S.C., Guest J.R. Biochem. J. 255:35-43(1988).</p>
Inos-1-P_synth		Myo-inositol-1-phosphate synthase	<p>Accession number: PF01658 Definition: Myo-inositol-1-phosphate synthase Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_959 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 86.80 86.80 Noise cutoffs: -219.00 -219.00 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 95066381 Reference Title: Comparison of INO1 gene sequences and products in <i>Candida albicans</i> and <i>Saccharomyces cerevisiae</i>. Reference Author: Klig LS, Zobel PA, Devry CG, Losberger C; Reference Location: Yeast 1994;10:789-800. Database Reference INTERPRO; IPR002587; Comment: This is a family of myo-inositol-1-phosphate synthases. Comment: Inositol-1-phosphate catalyses the conversion of glucose-6-phosphate to inositol-1-phosphate, which is then dephosphorylated to inositol [1]. Inositol phosphates play an important role in signal transduction. Comment: Number of members: 27</p>

IPP_isomerase		Isopentenyl-diphosphate delta-isomerase	<p>Accession number: PF01772</p> <p>Definition: Isopentenyl-diphosphate delta-isomerase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1099 (release 4.2)</p> <p>Gathering cutoffs: -88 -88</p> <p>Trusted cutoffs: -66.70 -66.70</p> <p>Noise cutoffs: -106.90 -106.90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98409684</p> <p>Reference Title: Differential expression of two isopentenyl pyrophosphate isomerases and enhanced carotenoid accumulation in a unicellular chlorophyte</p> <p>Reference Author: Sun Z, Cunningham FX Jr, Gantt E;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1998;95:11482-11488.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97373600</p> <p>Reference Title: Cloning and subcellular localization of hamster and rat isopentenyl diphosphate dimethylallyl diphosphate isomerase. A PTS1 motif targets the enzyme to peroxisomes.</p> <p>Reference Author: Paton VG, Shackelford JE, Krisans SK;</p> <p>Reference Location: J Biol Chem 1997;272:18945-18950.</p> <p>Database Reference INTERPRO; IPR002667;</p> <p>Comment: Isopentenyl-diphosphate delta-isomerase or IPP isomerase EC:5.3.3.2</p> <p>Comment: catalyses the interconversion of isopentenyl diphosphate and</p> <p>Comment: dimethylallyl diphosphate. Dimethylallyl phosphate is the initial substrate</p> <p>Comment: for the biosynthesis of carotenoids and other long chain isoprenoids [1].</p> <p>Number of members: 24</p>
K-box	PDOC00302	MADS-box domain signature and profile	<p>A number of transcription factors contain a conserved domain of 56 amino-acid residues, sometimes known as the MADS-box domain [E1]. They are listed below:</p> <ul style="list-style-type: none"> - Serum response factor (SRF) [1], a mammalian transcription factor that binds to the Serum Response Element (SRE). This is a short sequence of dyad symmetry located 300 bp to the 5' end of the transcription initiation site of genes such as c-fos. - Mammalian myocyte-specific enhancer factors 2A to 2D (MEF2A to MEF2D). <p>These proteins are transcription factor which binds specifically to the MEF2 element present in the regulatory regions of many muscle-specific genes.</p> <ul style="list-style-type: none"> - Drosophila myocyte-specific enhancer factor 2 (MEF2). - Yeast GRM/PRTF protein (gene MCM1) [2], a transcriptional regulator of mating-type-specific genes. - Yeast arginine metabolism regulation protein I (gene ARGR1 or ARG80). - Yeast transcription factor RLM1. - Yeast transcription factor SMP1. - Arabidopsis thaliana agamous protein (AG) [3], a probable transcription factor involved in regulating genes that determines stamen and carpel development in wild-type flowers. Mutations in the AG gene result in the replacement of the stamens by petals and the carpels by a new flower. - Arabidopsis thaliana homeotic proteins Apetala1 (AP1), Apetala3 (AP3) and Pistillata (PI) which act locally to specify the identity of the floral meristem and to determine sepal and petal development [4]. - Antirrhinum majus and tobacco homeotic protein deficiens (DEFA) and globosa (GLO) [5]. Both proteins are transcription factors involved in the genetic control of flower development. Mutations in DEFA or GLO cause the transformation of petals into sepals and of stamens into carpels. - Arabidopsis thaliana putative transcription factors AGL1 to AGL6 [6]. - Antirrhinum majus morphogenetic protein DEF H33 (squamosa). <p>In SRF, the conserved domain has been shown [1] to be involved in DNA-binding</p>

			<p>and dimerization. We have derived a pattern that spans the complete length of the domain. The profile also spans the length of the MADS-box.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern R-x-[RK]-x(5)-l-x-[DNGSK]-x(3)-[KR]-x(2)-T-[FY]-x-[RK](3)-x(2)-[LIVM]-x-K(2)-A-x-E-[LIVM]-[STA]-x-L-x(4)-[LIVM]-x-[LIVM](3)-x(6)-[LIVMF]-x(2)-[FY]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the profile ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this documentation entry is linked to both signature patterns and a profile. As the profile is much more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so.</p> <p>Last update July 1999 / Pattern and text revised.</p> <p>References</p> <p>[1] Norman C., Runswick M., Pollock R., Treisman R. Cell 55:989-1003(1988).</p> <p>[2] Passmore S., Maine G.T., Elble R., Christ C., Tye B.-K. J. Mol. Biol. 204:593-606(1988).</p> <p>[3] Yanofsky M., Ma H., Bowman J., Drews G., Feldmann K.A., Meyerowitz E.M. Nature 346:35-39(1990).</p> <p>[4] Goto K., Meyerowitz E.M. Genes Dev. 8:1548-1560(1994).</p> <p>[5] Troebner W., Ramirez L., Motte P., Hue I., Huijser P., Loennig W.-E., Saedler H., Sommer H., Schwartz-Sommer Z. EMBO J. 11:4693-4704(1992).</p> <p>[6] Ma H., Yanofsky M.F., Meyerowitz E.M. Genes Dev. 5:484-495(1991).</p> <p>[E1] http://transfac.gbf-braunschweig.de/cgi-bin/qt/getEntry.pl?C0014</p>
Keratin_B2		Keratin, high sulfur B2 protein	<p>Accession number: PF01500</p> <p>Definition: Keratin, high sulfur B2 protein</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_706 (release 4.0)</p> <p>Gathering cutoffs: -17 -17</p> <p>Trusted cutoffs: -1.50 -1.50</p> <p>Noise cutoffs: -46.00 18.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98201605</p> <p>Reference Title: Structure and hair follicle-specific expression of genes encoding the rat high sulfur protein B2 family.</p> <p>Reference Author: Mitsui S, Ohuchi A, Adachi-Yamada T, Hotta M, Tsuboi R,</p> <p>Reference Author: Ogawa H;</p> <p>Reference Location: Gene 1998;208:123-129.</p> <p>Database Reference INTERPRO: IPR002494;</p> <p>Comment: High sulfur proteins are cysteine-rich proteins synthesized during the differentiation of hair matrix cells, and form hair fibers in association with hair keratin intermediate filaments [1].</p> <p>Comment: This family has been divided up into four regions, with the</p>

			<p>second</p> <p>Comment: region containing 8 copies of a short repeat [1]. This family is</p> <p>Comment: also known as B2 or KAP1.</p> <p>Number of members: 17</p>
ketoacyl-synt	PDOC00529	Beta-ketoacyl synthases active site	<p>Beta-ketoacyl-ACP synthase (EC 2.3.1.41) (KAS) [1] is the enzyme that catalyzes the condensation of malonyl-ACP with the growing fatty acid chain. It is found as a component of the following enzymatic systems:</p> <ul style="list-style-type: none"> - Fatty acid synthetase (FAS), which catalyzes the formation of long-chain fatty acids from acetyl-CoA, malonyl-CoA and NADPH. Bacterial and plant chloroplast FAS are composed of eight separate subunits which correspond to different enzymatic activities; beta-ketoacyl synthase is one of these polypeptides. Fungal FAS consists of two multifunctional proteins, FAS1 and FAS2; the beta-ketoacyl synthase domain is located in the C-terminal section of FAS2. Vertebrate FAS consists of a single multifunctional chain; the beta-ketoacyl synthase domain is located in the N-terminal section [2]. - The multifunctional 6-methylsalicylic acid synthase (MSAS) from <i>Penicillium patulum</i> [3]. This is a multifunctional enzyme involved in the biosynthesis of a polyketide antibiotic and which has a KAS domain in its N-terminal section. - Polyketide antibiotic synthase enzyme systems. Polyketides are secondary metabolites produced by microorganisms and plants from simple fatty acids. KAS is one of the components involved in the biosynthesis of the <i>Streptomyces</i> polyketide antibiotics granatacin [4], tetracenomycin C [5] and erythromycin. - <i>Emericella nidulans</i> multifunctional protein Wa. Wa is involved in the biosynthesis of conidial green pigment. Wa is protein of 216 Kd that contains a KAS domain. - <i>Rhizobium</i> nodulation protein nodE, which probably acts as a beta-ketoacyl synthase in the synthesis of the nodulation Nod factor fatty acyl chain. - Yeast mitochondrial protein CEM1. <p>The condensation reaction is a two step process: the acyl component of an activated acyl primer is transferred to a cysteine residue of the enzyme and is then condensed with an activated malonyl donor with the concomitant release of carbon dioxide. The sequence around the active site cysteine is well conserved and can be used as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-x(4)-[LIVMFAP]-x(2)-[AGC]-C-[STA](2)-[STAG]-x(3)-[LIVMF] [C is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for bacterial and plant beta-ketoacyl synthase III (KAS III).</p> <p>Other sequence(s) detected in SWISS-PROT 10.</p> <p>Last update November 1997 / Text revised.</p> <p>References</p> <p>[1] Kauppinen S., Siggaard-Andersen M., von Wettstein-Knowles P. Carlsberg Res. Commun. 53:357-370(1988).</p> <p>[2] Witkowski A., Rangan V.S., Randhawa Z.I., Amy C.M., Smith S. Eur. J. Biochem. 198:571-579(1991).</p> <p>[3] Beck J., Ripka S., Siegner A., Schiltz E., Schweizer E. Eur. J. Biochem. 192:487-498(1990).</p> <p>[4] Bibb M.J., Biro S., Motamedi H., Collins J.F., Hutchinson C.R. EMBO J. 8:2727-2736(1989).</p> <p>[5] Sherman D.H., Malpartida F., Bibb M.J., Kieser H.M., Bibb M.J., Hopwood D.A. EMBO J. 8:2717-2725(1989).</p>
KRAB		KRAB box	<p>Accession number: PF01352</p> <p>Definition: KRAB box</p>

925

			<p>Author: Bateman A Alignment method of seed: Manual Source of seed members: Bateman A Gathering cutoffs: 0 0 Trusted cutoffs: 1.10 1.10 Noise cutoffs: -5.40 -5.40 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 91319563 Reference Title: Conserved KRAB protein domain identified upstream from the Reference Title: zinc finger region of Kox 8. Reference Author: Thiesen HJ, Bellefroid E, Revelant O, Martial JA; Reference Location: Nucleic Acids Res 1991;19:3996-3996. Reference Number: [2] Reference Medline: 97140325 Reference Title: A novel member of the RING finger family, KRIP-1, associates with the KRAB-A transcriptional repressor domain Reference Title: of zinc finger proteins. Reference Author: Kim SS, Chen YM, O'Leary E, Witzgall R, Vidal M, Bonventre Reference Author: JV; Reference Location: Proc Natl Acad Sci U S A 1996;93:15299-15304. Reference Number: [3] Reference Medline: 96365472 Reference Title: KAP-1, a novel corepressor for the highly conserved KRAB Reference Title: repression domain. Reference Author: Friedman JR, Fredericks WJ, Jensen DE, Speicher DW, Huang Reference Author: XP, Neilson EG, Rauscher FJ; Reference Location: Genes Dev 1996;10:2067-2078. Database Reference: INTERPRO; IPR001909; Database reference: PFAMB; PB036541; Comment: The KRAB domain (or Kruppel-associated box) is present in Comment: about a third of zinc finger proteins containing C2H2 fingers. Comment: The KRAB domain is found to be involved in protein-protein Comment: interactions [2,3]. Comment: The KRAB domain is generally encoded by two exons. The Comment: regions coded by the two exons are known as KRAB-A and Comment: KRAB-B. Number of members: 105</p>
lectin_legB	PDOC00278	Legume lectins signatures	<p>Leguminous plants synthesize sugar-binding proteins which are called legume lectins [1,2]. These lectins are generally found in the seeds. The exact function of legume lectins is not known but they may be involved in the attachment of nitrogen-fixing bacteria to legumes and in the protection against pathogens. Legume lectins bind calcium and manganese (or other transition metals).</p> <p>Legume lectins are synthesized as precursor proteins of about 230 to 260 amino acid residues. Some legume lectins are proteolytically processed to produce two chains: beta (which corresponds to the N-terminal) and alpha (C-terminal). The lectin concanavalin A (conA) from jack bean is exceptional in that the two chains are transposed and ligated (by formation of a new peptide bond). The N-terminus of mature conA thus corresponds to that of the alpha chain and the C-terminus to the beta chain.</p> <p>We have developed two signature patterns specific to legume lectins: the first is located in the C-terminal section of the beta chain and contains a conserved aspartic acid residue important for the binding of calcium and manganese; the second one is located in the N-terminal of the alpha chain.</p> <p>Description of pattern(s) and/or profile(s)</p>

			<p>Consensus pattern [LIV]-[STAG]-V-[DEQV]-[FLI]-D-[ST] [D binds manganese and calcium] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 21.</p> <p>Consensus pattern [LIV]-x-[EDQ]-[FYWKR]-V-x-[LIVF]-G-[LF]-[ST] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 4. Last update July 1999 / Patterns and text revised. References [1] Sharon N., Lis H. FASEB J. 4:3198-320(1990).</p> <p>[2] Lis H., Sharon N. Annu. Rev. Biochem. 55:33-37(1986).</p>
ligase-CoA		CoA-ligases	<p>Accession number: PF00549 Definition: CoA-ligases Author: Bateman A Alignment method of seed: Clustalw Source of seed members: SCOP Gathering cutoffs: 25 25 Trusted cutoffs: 28.70 28.70 Noise cutoffs: 14.70 14.70 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmlcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 94193797 Reference Title: The crystal structure of succinyl-CoA synthetase from Escherichia coli at 2.5-A resolution. Reference Author: Wolodko WT, Fraser ME, James MN, Bridger WA; Reference Location: J Biol Chem 1994;269:10883-10890. Database Reference: SCOP; 1scu; sf; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR000303; Database Reference: PDB; 1cqi A; 132; 279; Database Reference: PDB; 1cqi D; 132; 279; Database Reference: PDB; 1cqj A; 132; 279; Database Reference: PDB; 1cqj D; 132; 279; Database Reference: PDB; 2scu A; 132; 279; Database Reference: PDB; 2scu D; 132; 279; Database Reference: PDB; 1scu A; 132; 279; Database Reference: PDB; 1scu D; 132; 279; Database Reference: PDB; 1cqi B; 246; 385; Database Reference: PDB; 1cqi E; 246; 385; Database Reference: PDB; 1cqj B; 246; 385; Database Reference: PDB; 1cqj E; 246; 385; Database Reference: PDB; 2scu B; 246; 385; Database Reference: PDB; 2scu E; 246; 385; Database Reference: PDB; 1scu B; 246; 388; Database Reference: PDB; 1scu E; 246; 388; Database reference: PFAMB; PB039724; Database reference: PFAMB; PB041236; Comment: -!- This family includes the CoA ligases Succinyl-CoA synthetase alpha Comment: and beta chains, malate CoA ligase and ATP-citrate lyase. Comment: Some members of the family utilise ATP others use GTP. Number of members: 76</p>
LIM_bind		LIM-domain binding protein	<p>Accession number: PF01803 Definition: LIM-domain binding protein Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1352 (release 4.2) Gathering cutoffs: -92 -92 Trusted cutoffs: 13.40 13.40 Noise cutoffs: -197.90 -197.90 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmlcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97477378</p>

927

			<p>Reference Title: Chip, a widely expressed chromosomal protein required for</p> <p>Reference Title: segmentation and activity of a remote wing margin enhancer</p> <p>Reference Title: in Drosophila.</p> <p>Reference Author: Morcillo P, Rosen C, Baylies MK, Dorsett D;</p> <p>Reference Location: Genes Dev 1997;11:2729-2740.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97336071</p> <p>Reference Title: A family of LIM domain-associated cofactors confer transcriptional synergism between LIM and Otx homeodomain</p> <p>Reference Title: proteins.</p> <p>Reference Author: Bach I, Carriere C, Ostendorff HP, Andersen B, Rosenfeld</p> <p>Reference Author: MG;</p> <p>Reference Location: Genes Dev 1997;11:1370-1380.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 97078753</p> <p>Reference Title: Interactions of the LIM-domain-binding factor Ldb1 with LIM</p> <p>Reference Title: homeodomain proteins.</p> <p>Reference Author: Agulnick AD, Taira M, Breen JJ, Tanaka T, Dawid IB, Westphal H;</p> <p>Reference Location: Nature 1996;384:270-272.</p> <p>Reference Number: [4]</p> <p>Reference Medline: 97030257</p> <p>Reference Title: Nuclear LIM interactor, a rhombotin and LIM homeodomain</p> <p>Reference Title: interacting protein, is expressed early in neuronal development.</p> <p>Reference Title:</p> <p>Reference Author: Jurata LW, Kenny DA, Gill GN;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1996;93:11693-11698.</p> <p>Database Reference INTERPRO; IPR002691;</p> <p>Comment: The LIM-domain binding protein, binds to the LIM domain LIM of</p> <p>Comment: LIM homeodomain proteins which are transcriptional regulators of</p> <p>Comment: development.</p> <p>Comment: Nuclear LIM interactor (NLI) / LIM domain-binding protein 1 (LDB1)</p> <p>Comment: Swiss:P70662 is located in the nuclei of neuronal cells during</p> <p>Comment: development, it is co-expressed with Isl1 in early motor neuron</p> <p>Comment: differentiation and has a suggested role in the Isl1 dependent</p> <p>Comment: development of motor neurons [4].</p> <p>Comment: It is suggested that these proteins act synergistically to enhance</p> <p>Comment: transcriptional efficiency by acting as co-factors for LIM homeodomain</p> <p>Comment: and Otx class transcription factors both of which have essential roles</p> <p>Comment: in development [2].</p> <p>Comment: The Drosophila protein Chip Swiss:O18353 is required for segmentation</p> <p>Comment: and activity of a remote wing margin enhancer [1]. Chip is a ubiquitous</p> <p>Comment: chromosomal factor required for normal expression of diverse genes at</p> <p>Comment: many stages of development [1]. It is suggested that Chip cooperates</p> <p>Comment: with different LIM domain proteins and other factors to structurally</p> <p>Comment: support remote enhancer-promoter interactions [1].</p> <p>Number of members: 19</p>
Lipase_3	PDOC00110	Lipases, serine active site	<p>Triglyceride lipases (EC 3.1.1.3) [1] are lipolytic enzymes that hydrolyzes the ester bond of triglycerides. Lipases are widely distributed in animals, plants and prokaryotes. In higher vertebrates there are at least three tissue-specific isozymes: pancreatic, hepatic, and gastric/lingual. These three types of lipases are closely related to each other as well as to lipoprotein lipase (EC 3.1.1.34) [2], which hydrolyzes triglycerides of chylomicrons and very low</p>

928

			<p>density lipoproteins (VLDL).</p> <p>The most conserved region in all these proteins is centered around a serine residue which has been shown [3] to participate, with an histidine and an aspartic acid residue, to a charge relay system. Such a region is also present in lipases of prokaryotic origin and in lecithin-cholesterol acyltransferase (EC 2.3.1.43) (LCAT) [4], which catalyzes fatty acid transfer between phosphatidylcholine and cholesterol. We have built a pattern from that region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIV]-x-[LIVFY]-[LIVMST]-G-[HYWV]-S-x-G-[GSTAC] [S is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT 35.</p> <p>Note Drosophila vitellogenins are also related to lipases [5], but they have lost their active site serine.</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References</p> <p>[1] Chapus C., Rovey M., Sarda L., Verger R. Biochimie 70:1223-1234(1988).</p> <p>[2] Persson B., Bengtsson-Olivecrona G., Enerback S., Olivecrona T., Joernvall H. Eur. J. Biochem. 179:39-45(1989).</p> <p>[3] Blow D. Nature 343:694-695(1990).</p> <p>[4] McLean J., Fielding C., Drayna D., Dieplinger H., Baer B., Kohr W., Henzel W., Lawn R. Proc. Natl. Acad. Sci. U.S.A. 83:2335-2339(1986).</p> <p>[5] Baker M.E. Biochem. J. 255:1057-1060(1988).</p>
Lipase_GDSL	PDOC00842	Lipolytic enzymes "G-D-S-L" family, serine active site	<p>Recently [1], a family of lipolytic enzymes has been characterized. This family currently consist of the following proteins:</p> <ul style="list-style-type: none"> - Aeromonas hydrophila lipase/phosphatidylcholine-sterol acyltransferase. - Xenorhabdus luminescens lipase 1. - Vibrio mimicus arylesterase. - Escherichia coli acyl-coA thioesterase I (gene tesA). - Vibrio parahaemolyticus thermolabile hemolysin/atypical phospholipase. - Rabbit phospholipase AdRab-B, an intestinal brush border protein with esterase and phospholipase A/lysophospholipase activity that could be involved in the uptake of dietary lipids. AdRab-B contains four repeats of about 320 amino acids. - Arabidopsis thaliana and Brassica napus anther-specific proline-rich protein APG. - A Pseudomonas putida hypothetical protein in trpE-trpG intergenic region. <p>A serine has been identified a part of the active site in the Aeromonas, Vibrio mimicus and Escherichia coli enzymes. It is located in a conserved sequence motif that can be used as a signature pattern for these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMFYAG](4)-G-D-S-[LIVM]-x(1,2)-[TAG]-G [S is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this pattern will pick up two of the four repeats in AdRab-B, the first one is</p>

			<p>not detected as its sequence has diverged in the region of the putative active site residue. The last one is also not detected because it is slightly divergent at the end of the pattern.</p> <p>Expert(s) to contact by email Upton C. upton@sol.uvic.ca</p> <p>Buckley J.T. tbuckley@sol.uvic.ca</p> <p>Last update November 1995 / First entry.</p> <p>References [1] Upton C., Buckley J.T. <i>Trends Biochem. Sci.</i> 20:178-179(1995).</p>
Lipoprotein_1	PDOC00013	Prokaryotic membrane lipoprotein lipid attachment site	<p>In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):</p> <ul style="list-style-type: none"> - Major outer membrane lipoprotein (murein-lipoproteins) (gene <i>lpp</i>). - <i>Escherichia coli</i> lipoprotein-28 (gene <i>nlpA</i>). - <i>Escherichia coli</i> lipoprotein-34 (gene <i>nlpB</i>). - <i>Escherichia coli</i> lipoprotein <i>nlpC</i>. - <i>Escherichia coli</i> lipoprotein <i>nlpD</i>. - <i>Escherichia coli</i> osmotically inducible lipoprotein B (gene <i>osmB</i>). - <i>Escherichia coli</i> osmotically inducible lipoprotein E (gene <i>osmE</i>). - <i>Escherichia coli</i> peptidoglycan-associated lipoprotein (gene <i>pal</i>). - <i>Escherichia coli</i> rare lipoproteins A and B (genes <i>rplA</i> and <i>rplB</i>). - <i>Escherichia coli</i> copper homeostasis protein <i>cutF</i> (or <i>nlpE</i>). - <i>Escherichia coli</i> plasmids <i>traT</i> proteins. - <i>Escherichia coli</i> Col plasmids lysis proteins. - A number of <i>Bacillus</i> beta-lactamases. - <i>Bacillus subtilis</i> periplasmic oligopeptide-binding protein (gene <i>oppA</i>). - <i>Borrelia burgdorferi</i> outer surface proteins A and B (genes <i>ospA</i> and <i>ospB</i>). - <i>Borrelia hermsii</i> variable major protein 21 (gene <i>vmp21</i>) and 7 (gene <i>vmp7</i>). - <i>Chlamydia trachomatis</i> outer membrane protein 3 (gene <i>omp3</i>). - <i>Fibrobacter succinogenes</i> endoglucanase <i>cel-3</i>. - <i>Haemophilus influenzae</i> proteins <i>Pal</i> and <i>Pcp</i>. - <i>Klebsiella</i> pullulanase (gene <i>pulA</i>). - <i>Klebsiella</i> pullulanase secretion protein <i>pulS</i>. - <i>Mycoplasma hyorhinis</i> protein <i>p37</i>. - <i>Mycoplasma hyorhinis</i> variant surface antigens A, B, and C (genes <i>vipABC</i>). - <i>Neisseria</i> outer membrane protein H.8. - <i>Pseudomonas aeruginosa</i> lipopeptide (gene <i>lppL</i>). - <i>Pseudomonas solanacearum</i> endoglucanase <i>egl</i>. - <i>Rhodopseudomonas viridis</i> reaction center cytochrome subunit (gene <i>cytC</i>). - <i>Rickettsia</i> 17 Kd antigen. - <i>Shigella flexneri</i> invasion plasmid proteins <i>mxiJ</i> and <i>mxiM</i>. - <i>Streptococcus pneumoniae</i> oligopeptide transport protein A (gene <i>amiA</i>). - <i>Treponema pallidum</i> 34 Kd antigen. - <i>Treponema pallidum</i> membrane protein A (gene <i>tmpA</i>). - <i>Vibrio harveyi</i> chitinase (gene <i>chb</i>). - <i>Yersinia</i> virulence plasmid protein <i>yscJ</i>. <p>- Halocyanin from <i>Natrobacterium pharaonis</i> [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion).</p> <p>From the precursor sequences of all these proteins, we derived a consensus pattern and a set of rules to identify this type of post-translational modification.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern {DERK}(6)-[LIVFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least</p>

			<p>one Lys or one Arg in the first seven positions of the sequence. Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be. Last update November 1995 / Pattern and text revised. References</p> <p>[1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).</p> <p>[2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).</p> <p>[3] von Heijne G. Protein Eng. 2:531-534(1989).</p> <p>[4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterheld D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).</p>
Lipoprotein_2	PDOC00013	Prokaryotic membrane lipoprotein lipid attachment site	<p>In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):</p> <ul style="list-style-type: none"> - Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp). - Escherichia coli lipoprotein-28 (gene nlpA). - Escherichia coli lipoprotein-34 (gene nlpB). - Escherichia coli lipoprotein nlpC. - Escherichia coli lipoprotein nlpD. - Escherichia coli osmotically inducible lipoprotein B (gene osmB). - Escherichia coli osmotically inducible lipoprotein E (gene osmE). - Escherichia coli peptidoglycan-associated lipoprotein (gene pal). - Escherichia coli rare lipoproteins A and B (genes rplA and rplB). - Escherichia coli copper homeostasis protein cutF (or nlpE). - Escherichia coli plasmids traT proteins. - Escherichia coli Col plasmids lysis proteins. - A number of Bacillus beta-lactamases. - Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA). - Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB). - Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7). - Chlamydia trachomatis outer membrane protein 3 (gene omp3). - Fibrobacter succinogenes endoglucanase cel-3. - Haemophilus influenzae proteins Pal and Pcp. - Klebsiella pullulunase (gene pulA). - Klebsiella pullulunase secretion protein pulS. - Mycoplasma hyorhinis protein p37. - Mycoplasma hyorhinis variant surface antigens A, B, and C (genes vlpABC). - Neisseria outer membrane protein H.8. - Pseudomonas aeruginosa lipopeptide (gene lppL). - Pseudomonas solanacearum endoglucanase egl. - Rhodopseudomonas viridis reaction center cytochrome subunit (gene cytC). - Rickettsia 17 Kd antigen. - Shigella flexneri invasion plasmid proteins mxj and mxiM. - Streptococcus pneumoniae oligopeptide transport protein A (gene amiA). - Treponema pallidum 34 Kd antigen. - Treponema pallidum membrane protein A (gene tmpA). - Vibrio harveyi chitobiase (gene chb). - Yersinia virulence plasmid protein yscJ. <p>- Halocyanin from Natrobacterium pharaonis [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion).</p> <p>From the precursor sequences of all these proteins, we derived a consensus pattern and a set of rules to identify this type of post-translational modification.</p>

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence. Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be. Last update November 1995 / Pattern and text revised. References [1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990). [2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988). [3] von Heijne G. Protein Eng. 2:531-534(1989). [4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).</p>
Lipoprotein_5	PDOC00013	Prokaryotic membrane lipoprotein lipid attachment site	<p>In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):</p> <ul style="list-style-type: none"> - Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp). - Escherichia coli lipoprotein-28 (gene nlpA). - Escherichia coli lipoprotein-34 (gene nlpB). - Escherichia coli lipoprotein nlpC. - Escherichia coli lipoprotein nlpD. - Escherichia coli osmotically inducible lipoprotein B (gene osmB). - Escherichia coli osmotically inducible lipoprotein E (gene osmE). - Escherichia coli peptidoglycan-associated lipoprotein (gene pal). - Escherichia coli rare lipoproteins A and B (genes rplA and rplB). - Escherichia coli copper homeostasis protein cutF (or nlpE). - Escherichia coli plasmids traT proteins. - Escherichia coli Col plasmids lysis proteins. - A number of Bacillus beta-lactamases. - Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA). - Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB). - Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7). - Chlamydia trachomatis outer membrane protein 3 (gene omp3). - Fibrobacter succinogenes endoglucanase cel-3. - Haemophilus influenzae proteins Pal and Pcp. - Klebsiella pullulunase (gene pulA). - Klebsiella pullulunase secretion protein pulS. - Mycoplasma hyorhinis protein p37. - Mycoplasma hyorhinis variant surface antigens A, B, and C (genes vlpABC). - Neisseria outer membrane protein H.8. - Pseudomonas aeruginosa lipopeptide (gene lppL). - Pseudomonas solanacearum endoglucanase egl. - Rhodopseudomonas viridis reaction center cytochrome subunit (gene cytC). - Rickettsia 17 Kd antigen. - Shigella flexneri invasion plasmid proteins mxiJ and mxiM. - Streptococcus pneumoniae oligopeptide transport protein A (gene amiA). - Treponema pallidum 34 Kd antigen. - Treponema pallidum membrane protein A (gene tmpA). - Vibrio harveyi chitinase (gene chb). - Yersinia virulence plasmid protein yscJ.

			<p>- Halocyanin from <i>Natrobacterium pharaonis</i> [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion).</p> <p>From the precursor sequences of all these proteins, we derived a consensus pattern and a set of rules to identify this type of post-translational modification.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern {DERK}(6)-[LIVFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence.</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be.</p> <p>Last update November 1995 / Pattern and text revised.</p> <p>References [1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).</p> <p>[2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).</p> <p>[3] von Heijne G. Protein Eng. 2:531-534(1989).</p> <p>[4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).</p>
Luteo_Vpg		Luteovirus putative VPg genome linked protein	<p>Accession number: . PF01659</p> <p>Definition: Luteovirus putative VPg genome linked protein</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_970 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 191.70 191.70</p> <p>Noise cutoffs: -47.90 -47.90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 94120742</p> <p>Reference Title: Soybean dwarf luteovirus contains the third variant genome</p> <p>Reference Title: type in the luteovirus group.</p> <p>Reference Author: Rathjen JP, Karageorgos LE, Habili N, Waterhouse PM, Symons</p> <p>Reference Author: RH;</p> <p>Reference Location: Virology 1994;198:671-679.</p> <p>Database Reference: INTERPRO; IPR001964;</p> <p>Comment: This family consists of several putative genome linked proteins.</p> <p>Comment: The genomic RNA of luteoviruses are linked to virally encoded genome</p> <p>Comment: proteins (VPg). Open reading frame 4 is thought to encode the VPg</p> <p>Comment: in Soybean dwarf luteovirus [1].</p> <p>Comment: Luteoviruses have isometric capsids that contain a positive strand</p> <p>Comment: ssRNA genome, they have no DNA stage during their replication.</p> <p>Number of members: 32</p>
MATH		MATH domain	<p>Accession number: PF00917</p> <p>Definition: MATH domain</p>

933

			<p>Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1602 (release 3.0) Gathering cutoffs: 17 0 Trusted cutoffs: 17.90 0.20 Noise cutoffs: 11.80 11.80 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmlcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96334294 Reference Title: TRAF proteins and meprins share a conserved domain. Reference Author: Uren AG, Vaux DL; Reference Location: Trends Biochem Sci 1996;21:244-245. Reference Number: [2] Reference Medline: 99342031 Reference Title: Crystallographic analysis of CD40 recognition and signaling Reference Title: by human TRAF2. Reference Author: McWhirter SM, Pullen SS, Holton JM, Crute JJ, Kehry MR, Reference Author: Alber T; Reference Location: Proc Natl Acad Sci U S A 1999;96:8408-8413. Reference Number: [3] Reference Medline: 99069615 Reference Title: Comparison of the complete protein sets of worm and yeast: Reference Title: orthology and divergence. Reference Author: Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Reference Author: Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, Weng S, Reference Author: Cherry JM, Botstein D; Reference Location: Science 1998;282:2022-2028. Database Reference: SCOP; 1qsc; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002083; Database Reference: PDB; 1qsc A; 357; 498; Database Reference: PDB; 1qsc B; 357; 498; Database Reference: PDB; 1qsc C; 357; 498; Database reference: PFAMB; PB018448; Database reference: PFAMB; PB040690; Database reference: PFAMB; PB041198; Comment: This motif has been called the Meprin And TRAF-Homology Comment: (MATH) domain. This domain is hugely expanded in the nematode Comment: C. elegans [3]. Number of members: 212</p>
MCT		Monocarboxylate transporter	<p>Accession number: PF01587 Definition: Monocarboxylate transporter Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_483 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 322.90 322.90 Noise cutoffs: -38.20 -38.20 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmlcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98087501 Reference Title: Cloning and sequencing of four new mammalian monocarboxylate transporter (MCT) homologues confirms the Reference Title: existence of a transporter family with an ancient past. Reference Author: Price NT, Jackson VN, Halestrap AP, Reference Location: Biochem J 1998;329:321-328. Database Reference: INTERPRO; IPR002897; Comment: This domain consists of the transmembrane region of the monocarboxylate Comment: transporters. Monocarboxylate transporters (MTC) are transmembrane Comment: glycoproteins with 10-12 predicted transmembrane regions. Comment: They catalyse the proton linked transport of lactic acid,</p>

934

			<p>Comment: pyruvate and ketone bodies across the plasma membrane [1].</p> <p>Number of members: 33</p>
Methionine_synt		Methionine synthase, vitamin-B12 independent	<p>Accession number: PF01717</p> <p>Definition: Methionine synthase, vitamin-B12 independent</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1909 (release 4.1)</p> <p>Gathering cutoffs: -155.0 -155.0</p> <p>Trusted cutoffs: -155.00 -155.00</p> <p>Noise cutoffs: -170.00 -170.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98301657</p> <p>Reference Title: The specific features of methionine biosynthesis and metabolism in plants.</p> <p>Reference Author: Ravanel S, Gakiere B, Job D, Douce R;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1998;95:7805-7812.</p> <p>Database Reference: INTERPRO; IPR002629;</p> <p>Database reference: PFAMB; PB041617;</p> <p>Comment: This is a family of vitamin-B12 independent methionine synthases</p> <p>Comment: or 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferases, EC:2.1.1.14 from bacteria and plants.</p> <p>Comment: Plants are the only higher eukaryotes that have the required enzymes</p> <p>Comment: for methionine synthesis [1].</p> <p>Comment: This enzyme catalyses the last step in the production of methionine</p> <p>Comment: by transferring a methyl group from 5-methyltetrahydrofolate to homocysteine [1].</p> <p>Comment: The aligned region makes up the carboxy region of the approximately 750 amino acid protein except in some hypothetical archaeal proteins</p> <p>Comment: present in the family, where this region corresponds to the entire length.</p> <p>Number of members: 28</p>
Methyltransf_2		O-methyltransferase	<p>Accession number: PF00891</p> <p>Definition: O-methyltransferase</p> <p>Previous Pfam IDs: Methyltransf;</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_152 (release 3.0)</p> <p>Gathering cutoffs: -53 -53</p> <p>Trusted cutoffs: -22.00 -22.00</p> <p>Noise cutoffs: -84.60 -84.60</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 93167811</p> <p>Reference Title: Purification of a 40-kilodalton methyltransferase active in the aflatoxin biosynthetic pathway.</p> <p>Reference Author: Keller NP, Dischinger HC, Bhatnagar D, Cleveland TE, Ullah</p> <p>Reference Author: AH;</p> <p>Reference Location: Appl Environ Microbiol 1993;59:479-484.</p> <p>Database Reference: INTERPRO; IPR001077;</p> <p>Comment: This family includes a range of O-methyltransferases. These enzymes utilise S-adenosyl methionine.</p> <p>Number of members: 67</p>
Methyltransf_3		O-methyltransferase	<p>Accession number: PF01596</p> <p>Definition: O-methyltransferase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_749 (release 4.1)</p> <p>Gathering cutoffs: -86 -86</p> <p>Trusted cutoffs: -81.80 -81.80</p>

935

			<p>Noise cutoffs: -91.00 -91.00 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 97090395 Reference Title: Two multifunctional peptide synthetases and an Reference Title: O-methyltransferase are involved in the biosynthesis of the Reference Title: DNA-binding antibiotic and antitumour agent saframycin Mx1 Reference Title: from Myxococcus xanthus. Reference Author: Pospiech A, Bietenhader J, Schupp T; Reference Location: Microbiology 1996;142:741-746. Database Reference: SCOP; 1vid; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002935; Database Reference: PDB; 1vid; 13; 186; Database reference: PFAMB; PB040269; Comment: Members of this family are O-methyltransferases. The family Comment: includes catechol o-methyltransferase Swiss:P21964, caffeoyl-CoA Comment: O-methyltransferase Swiss:Q43095 and a family of bacterial Comment: O-methyltransferases that may be involved in antibiotic production [1]. Comment: Number of members: 39</p>
MMR_HSR1		GTPase of unknown function	<p>Accession number: PF01926 Definition: GTPase of unknown function Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: -21 -21 Trusted cutoffs: -20.70 -20.70 Noise cutoffs: -31.60 -31.60 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 94235953 Reference Title: Structure and evolution of a member of a new subfamily of Reference Title: GTP-binding proteins mapping to the human MHC class I region. Reference Title: Reference Author: Vernet C, Ribouchon MT, Chimini GPontarotti P; Reference Location: Mamm Genome 1994;5:100-105. Database Reference: INTERPRO; IPR002917; Database reference: PFAMB; PB000471; Database reference: PFAMB; PB002171; Database reference: PFAMB; PB015790; Number of members: 67</p>
MoaC		MoaC family	<p>Accession number: PF01967 Definition: MoaC family Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 73.00 73.00 Noise cutoffs: -93.90 -93.90 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 99337076 Reference Title: Characterization of a molybdenum cofactor biosynthetic gene Reference Title: cluster in Rhodobacter capsulatus which is specific for the Reference Title: biogenesis of dimethylsulfoxide reductase. Reference Author: Solomon PS, Shaw AL, Lane I, Hanson GR, Palmer T, McEwan Reference Author: AG; Reference Location: Microbiology 1999;145:1421-1429. Database Reference: INTERPRO; IPR002820; Comment: Members of this family are involved in molybdenum Comment: cofactor biosynthesis. However their molecular</p>

			<p>Comment: function is not known.</p> <p>Number of members: 24</p>
Myc_N_term		Myc amino-terminal region	<p>Accession number: PF01056</p> <p>Definition: Myc amino-terminal region</p> <p>Author: Finn RD, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_387 (release 3.0)</p> <p>Gathering cutoffs: -109 -109</p> <p>Trusted cutoffs: -81.20 -81.20</p> <p>Noise cutoffs: -137.40 -137.40</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98280742</p> <p>Reference Title: The molecular role of Myc in growth and transformation: recent discoveries lead to new insights.</p> <p>Reference Author: Facchini LM, Penn LZ;</p> <p>Reference Location: FASEB J 1998;12:633-651.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97318600</p> <p>Reference Title: Myc target genes.</p> <p>Reference Author: Grandori C, Eisenman RN;</p> <p>Reference Location: Trends Biochem Sci 1997;22:177-181.</p> <p>Database Reference INTERPRO; IPR002418;</p> <p>Comment: The myc family belongs to the basic helix-loop-helix leucine zipper</p> <p>Comment: class of transcription factors, see HLH. Myc forms a heterodimer with Max, and this complex regulates cell growth through</p> <p>Comment: direct activation of genes involved in cell replication [2].</p> <p>Number of members: 56</p>
Myosin_tail		Myosin tail	<p>Accession number: PF01576</p> <p>Definition: Myosin tail</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_356 (release 4.1)</p> <p>Gathering cutoffs: 19 19</p> <p>Trusted cutoffs: 23.30 23.30</p> <p>Noise cutoffs: 15.10 15.10</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 87060988</p> <p>Reference Title: Complete nucleotide and encoded amino acid sequence of a</p> <p>Reference Title: mammalian myosin heavy chain gene. Evidence against intron-dependent evolution of the rod.</p> <p>Reference Author: Strehler EE, Strehler-page M-A, Perriard JC, Periasamy M,</p> <p>Reference Author: Nadal-ginard B;</p> <p>Reference Location: J MOL BIOL 1986;190:291-317.</p> <p>Database Reference INTERPRO; IPR002928;</p> <p>Comment: The myosin molecule is a multi-subunit complex made up of two heavy chains and four light chains it is a fundamental contractile</p> <p>Comment: protein found in all eukaryote cell types [1].</p> <p>Comment: This family consists of the coiled-coil myosin heavy chain tail region.</p> <p>Comment: The coiled-coil is composed of the tail from two molecules of myosin.</p> <p>Comment: These can then assemble into the macromolecular thick filament [1].</p> <p>Comment: The coiled-coil region provides the structural backbone the thick</p> <p>Comment: filament [1].</p> <p>Number of members: 182</p>
Na_Ala_symp	PDOC00681	Sodium:alanine symporter family signature	<p>It has been shown [1] that integral membrane proteins that mediate the intake of a wide variety of molecules with the concomitant uptake of sodium ions (sodium symporters) can be grouped, on the basis of sequence and functional similarities into a number of distinct families. One of these families is known as the sodium:alanine symporter family (SAF) and currently consists of</p>

			<p>the following proteins:</p> <ul style="list-style-type: none"> - Thermophilic bacterium PS-3 alanine carrier protein (ACP). ACP can use both sodium and hydrogen as a symport ion. - Alteromonas haloplanktis D-alanine/glycine permease (gene dagA). - Bacillus subtilis alsT. - Hypothetical protein yaaJ from Escherichia coli and HI0183, the corresponding Haemophilus influenzae protein. - Haemophilus influenzae hypothetical protein HI0883. <p>These integral membrane proteins are predicted to comprise a least eight membrane spanning domains. As a signature pattern we selected a highly conserved region which is located in the N-terminal section and which includes part of the first transmembrane region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-G-x-[GA](2)-[LIVM]-F-W-M-W-[LIVM]-x-[STAV]-[LIVMFA](2)-G</p> <p>Sequences known to belong to this class detected by ône pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Reizer J., Reizer A., Saier M.H. Jr. Biochim. Biophys. Acta 1197:133-136(1994).</p>
Na_Ca_Ex		Sodium/calcium exchanger protein	<p>Accession number: PF01699</p> <p>Definition: Sodium/calcium exchanger protein</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1680 (release 4.1)</p> <p>Gathering cutoffs: 3 3</p> <p>Trusted cutoffs: 3.40 3.40</p> <p>Noise cutoffs: 1.20 1.20</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96394663</p> <p>Reference Title: Cloning of a third mammalian Na⁺-Ca²⁺ exchanger, NCX3.</p> <p>Reference Author: Nicoll DA, Quednau BD, Qui Z, Xia YR, Lusis AJ, Philipson</p> <p>Reference Author: KD;</p> <p>Reference Location: J Biol Chem 1996;271:24914-24921.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 91047958</p> <p>Reference Title: Molecular cloning and functional expression of the cardiac sarcolemmal Na(+)-Ca²⁺ exchanger.</p> <p>Reference Author: Nicoll DA, Longoni S, Philipson KD;</p> <p>Reference Location: Science 1990;250:562-565.</p> <p>Database Reference INTERPRO; IPR002613;</p> <p>Database reference: PFAMB; PB002768;</p> <p>Database reference: PFAMB; PB040773;</p> <p>Database reference: PFAMB; PB041540;</p> <p>Comment: This is a family of sodium/calcium exchanger integral membrane proteins. This family covers the integral membrane regions of the proteins. Sodium/calcium exchangers regulate intracellular Ca²⁺ concentrations in many cells; cardiac myocytes, epithelial cells, neurons retinal rod photoreceptors and smooth muscle cells [2].</p> <p>Comment: Ca²⁺ is moved into or out of the cytosol depending on Na⁺ concentration</p> <p>Comment: [2]. In humans and rats there are 3 isoforms; NCX1 NCX2 and NCX3 [1]</p> <p>Comment: see Swiss:Q01728, Swiss:P48768 and Swiss:P70549</p>

938

			respectively. Number of members: 105
Na_K_ATPas_e_C		Na+/K+ ATPase C-terminus	<p>This domain is specific to the sodium and potassium ATPases (Na_K-ATPase). The sodium pump (Na⁺/K⁺ ATPase), located in the plasma membrane of all animal cells [1], is an heterotrimer of a catalytic subunit (alpha chain), a glycoprotein subunit of about 34 Kd (beta chain) and a small hydrophobic protein of about 6 Kd. The beta subunit seems [2] to regulate, through the assembly of alpha/beta heterodimers, the number of sodium pumps transported to the plasma membrane.</p> <p>This family is typically found in association with E1-E2 ATPase. Uses of these polypeptide includes regulating that ion content in a desired cell or organism and can convey salt or ion tolerance.</p>
Na_K_ATPas_e_N		Na+/K+ ATPase C-terminus	<p>Accession number: PF00690 Definition: Na⁺/K⁺ ATPase C-terminus Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_138 (release 2.1) Gathering cutoffs: 15.6 15.6 Trusted cutoffs: 15.60 15.60 Noise cutoffs: 15.10 15.10 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Database Reference INTERPRO; IPR000661; Database reference: PFAMB; PB000031; Comment: This family is always found in association with E1-E2 ATPase. Comment: This extension is specific to the Na⁺/K⁺ ATPase subfamily of Comment: ATPases. Number of members: 90</p>
NAD_Gly3P_dh	PDOC00740	NAD-dependent glycerol-3-phosphate dehydrogenase signature	<p>NAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.1.8) (GPD) catalyzes the reversible reduction of dihydroxyacetone phosphate to glycerol-3-phosphate. It is a eukaryotic cytosolic homodimeric protein of about 40 Kd. As a signature pattern we selected a glycine-rich region that is probably [1] involved in NAD-binding.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-[AT]-[LIVM]-K-[DN]-[LIVM](2)-A-x-[GA]-x-G-[LIVMF]-x-[DE]-G-[LIVM]-x-[LIVMFYW]-G-x-N Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update November 1997 / Pattern and text revised. References [1] Otto J., Argos P., Rossmann M.G. Eur. J. Biochem. 109:325-330(1980).</p>
NifU_N		NifU-like N terminal domain	<p>Accession number: PF01592 Definition: NifU-like N terminal domain Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_772 (release 4.1) Gathering cutoffs: -13 -13 Trusted cutoffs: 1.20 1.20 Noise cutoffs: -28.80 -28.80 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97032601 Reference Title: A modular domain of NifU, a nitrogen fixation cluster protein, is highly conserved in evolution. Reference Author: Hwang DM, Dempsey A, Tan KT, Liew CC; Reference Location: J Mol Evol 1996;43:536-540. Database Reference INTERPRO; IPR002871; Comment: This domain is found in NifU in combination with NifU-like. Comment: This domain is found on isolated in several bacterial</p>

			<p>species</p> <p>Comment: such as Swiss:O53156. The nif genes are responsible for nitrogen fixation. However this domain is found in bacteria that do not fix nitrogen, so it may have a broader significance in the cell</p> <p>Comment: than nitrogen fixation.</p> <p>Number of members: 32</p>
NTR		NTR/C345C module	<p>Accession number: PF01759</p> <p>Definition: NTR/C345C module</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 57.30 57.30</p> <p>Noise cutoffs: 2.80 2.80</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 99379676</p> <p>Reference Title: The NTR module: domains of netrins, secreted frizzled related proteins, and type I procollagen C-proteinase enhancer protein are homologous with tissue inhibitors of metalloproteases [In Process Citation]</p> <p>Reference Author: Banyai L, Pathy L;</p> <p>Reference Location: Protein Sci 1999;8:1636-1642.</p> <p>Database Reference: INTERPRO: IPR001134;</p> <p>Database reference: PFAMB; PB005955;</p> <p>Comment: We have not included the related TIMP family.</p> <p>Comment: It has been suggested that the common function of these modules is binding to metzincins [1]. A subset of this family</p> <p>Comment: is known as the C345C domain because it occurs in complement</p> <p>Comment: C3, C4 and C5.</p> <p>Number of members: 64</p>
Nucleoside_transporter		Nucleoside transporter	<p>Accession number: PF01733</p> <p>Definition: Nucleoside transporter</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_2135 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 25.50 25.50</p> <p>Noise cutoffs: -122.50 -122.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98148080</p> <p>Reference Title: Cloning of the human equilibrative, nitrobenzylmercaptapurine riboside (NBMPR)-insensitive nucleoside transporter ei by functional expression in a transport-deficient cell line.</p> <p>Reference Author: Crawford CR, Patel DH, Naeve C, Belt JA;</p> <p>Reference Location: J Biol Chem 1998;273:5288-5293.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98019212</p> <p>Reference Title: Molecular cloning and functional characterization of nitrobenzylthioinosine (NBMPR)-sensitive (es) and NBMPR-insensitive (ei) equilibrative nucleoside transporter</p> <p>Reference Title: proteins (rENT1 and rENT2) from rat tissues.</p> <p>Reference Author: Yao SY, Ng AM, Muzyka WR, Griffiths M, Cass CE, Baldwin SA,</p> <p>Reference Author: Young JD;</p> <p>Reference Location: J Biol Chem 1997;272:28423-28430.</p> <p>Database Reference: INTERPRO: IPR002259;</p> <p>Comment: This is a family of nucleoside transporters.</p> <p>Comment: In mammalian cells nucleoside transporters transport nucleoside across the plasma membrane and are essential for nucleotide</p>

940

			<p>Comment: synthesis via the salvage pathways for cells that lack their own</p> <p>Comment: de novo synthesis pathways [2].</p> <p>Comment: Also in this family is mouse and human nucleolar protein HNP36</p> <p>Comment: Swiss:Q14542 a protein of unknown function; although it has been</p> <p>Comment: hypothesized to be a plasma membrane nucleoside transporter [2].</p> <p>Number of members: 15</p>
Orbi_VP6		Orbivirus helicase VP6	<p>Accession number: PF01516</p> <p>Definition: Orbivirus helicase VP6</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_765 (release 4.0)</p> <p>Gathering cutoffs: -68 -68</p> <p>Trusted cutoffs: -37.10 -37.10</p> <p>Noise cutoffs: -98.90 -98.90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97456481</p> <p>Reference Title: Bluetongue virus VP6 protein binds ATP and exhibits an RNA-dependent ATPase function and a helicase activity that</p> <p>Reference Title: catalyze the unwinding of double-stranded RNA substrates.</p> <p>Reference Author: Stauber N, Martinez-Costas J, Sutton G, Monastyrskaya K,</p> <p>Reference Author: Roy P;</p> <p>Reference Location: J Virol 1997;71:7220-7226.</p> <p>Database Reference: INTERPRO; IPR001399;</p> <p>Comment: The VP6 protein a minor protein in the core of the virion is probably the viral helicase [1].</p> <p>Number of members: 27</p>
OSCP	PDOC00327	ATP synthase delta (OSCP) subunit signature	<p>ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), which acts as a proton channel, and a catalytic core, termed coupling factor CF(1).</p> <p>One of the subunits of the ATPase complex, known as subunit delta in bacteria and chloroplasts or the Oligomycin Sensitivity Conferral Protein (OSCP) in mitochondria, seems to be part of the stalk that links CF(0) to CF(1). It either transmits conformational changes from CF(0) into CF(1) or is involved in proton conduction [3].</p> <p>The different delta/OSCP subunits are proteins of approximately 200 amino-acid residues - once the transit peptide has been removed in the chloroplast and mitochondrial forms - which show only moderate sequence homology.</p> <p>The signature pattern used to detect ATPase delta/OSCP subunits is based on a conserved region in the C-terminal section of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-x-[LIVMFYT]-x(3)-[LIVMT]-[DENQK]-x(2)-[LIVM]-x-[GSA]-G-[LIVMFYGA]-x-[LIVM]-[KRHENQ]-x-[GSEN]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except 3 sequences.</p> <p>Other sequence(s) detected in SWISS-PROT 2.</p> <p>Last update</p> <p>November 1997 / Pattern and text revised.</p> <p>References</p> <p>[1]</p> <p>Futai M., Noumi T., Maeda M.</p>

941

			<p>Annu. Rev. Biochem. 58:111-136(1989).</p> <p>[2] Senior A.E. Physiol. Rev. 68:177-231(1988).</p> <p>[3] Engelbrecht S., Junge W. Biochim. Biophys. Acta 1015:379-390(1990).</p>
OTCase	PDOC00091	Aspartate and ornithine carbamoyltransferases signature	<p>Aspartate carbamoyltransferase (EC 2.1.3.2) (ATCase) catalyzes the conversion of aspartate and carbamoyl phosphate to carbamoylaspartate, the second step in the de novo biosynthesis of pyrimidine nucleotides [1]. In prokaryotes ATCase consists of two subunits: a catalytic chain (gene pyrB) and a regulatory chain (gene pyrI), while in eukaryotes it is a domain in a multifunctional enzyme (called URA2 in yeast, rudimentary in Drosophila, and CAD in mammals [2]) that also catalyzes other steps of the biosynthesis of pyrimidines.</p> <p>Ornithine carbamoyltransferase (EC 2.1.3.3) (OTCase) catalyzes the conversion of ornithine and carbamoyl phosphate to citrulline. In mammals this enzyme participates in the urea cycle [3] and is located in the mitochondrial matrix. In prokaryotes and eukaryotic microorganisms it is involved in the biosynthesis of arginine. In some bacterial species it is also involved in the degradation of arginine [4] (the arginine deaminase pathway).</p> <p>It has been shown [5] that these two enzymes are evolutionary related. The predicted secondary structure of both enzymes are similar and there are some regions of sequence similarities. One of these regions includes three residues which have been shown, by crystallographic studies [6], to be implicated in binding the phosphoryl group of carbamoyl phosphate. We have selected this region as a signature for these enzymes.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern F-x-[EK]-x-S-[GT]-R-T [S, R, and the 2nd T bind carbamoyl phosphate] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note the residue in position 3 of the pattern allows to distinguish between an ATCase (Glu) and an OTCase (Lys). Last update October 1993 / Text revised. References [1] Lerner C.G., Switzer R.L. J. Biol. Chem. 261:11156-11165(1986).</p> <p>[2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B. BioEssays 15:157-164(1993).</p> <p>[3] Takiguchi M., Matsubasa T., Amaya Y., Mori M. BioEssays 10:163-166(1989).</p> <p>[4] Baur H., Stalon V., Falmagne P., Luethi E., Haas D. Eur. J. Biochem. 166:111-117(1987).</p> <p>[5] Houghton J.E., Bencini D.A., O'Donovan G.A., Wild J.R. Proc. Natl. Acad. Sci. U.S.A. 81:4864-4868(1981).</p> <p>[6] Ke H.-M., Honzatko R.B., Lipscomb W.N. Proc. Natl. Acad. Sci. U.S.A. 81:4037-4040(1984).</p>

942

oxidored_q1_N		NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminus	<p>Accession number: PF00662</p> <p>Definition: NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminus</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_22 (release 2.1)</p> <p>Gathering cutoffs: 18 18</p> <p>Trusted cutoffs: 19.40 19.40</p> <p>Noise cutoffs: 16.70 16.70</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmbuild --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 93110040</p> <p>Reference Title: The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains.</p> <p>Reference Author: Walker JE;</p> <p>Reference Location: Q Rev Biophys 1992;25:253-324.</p> <p>Database Reference: INTERPRO; IPR001516;</p> <p>Database reference: PFAMB; PB000410;</p> <p>Database reference: PFAMB; PB033295;</p> <p>Database reference: PFAMB; PB040550;</p> <p>Comment: This sub-family represents an amino terminal extension of oxidored_q1. Only NADH-Ubiquinone chain 5 and eubacterial chain L are in this family.</p> <p>Comment: This sub-family is part of complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane.</p> <p>Number of members: 546</p>
oxidored_q2		NADH-ubiquinone/plastoquinone oxidoreductase chain 4L	<p>Accession number: PF00420</p> <p>Definition: NADH-ubiquinone/plastoquinone oxidoreductase chain 4L</p> <p>Author: Finn RD</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_193 (release 1.0)</p> <p>Gathering cutoffs: 25 15</p> <p>Trusted cutoffs: 29.70 29.70</p> <p>Noise cutoffs: 20.40 20.40</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmbuild --seed 0 HMM</p> <p>Database Reference: INTERPRO; IPR001133;</p> <p>Database reference: PFAMB; PB006066;</p> <p>Number of members: 219</p>
PAN	PDOC00376	Apple domain	<p>Plasma kallikrein (EC 3.4.21.34) and coagulation factor XI (EC 3.4.21.27) are two related plasma serine proteases activated by factor XIIA and which share the same domain topology: an N-terminal region that contains four tandem repeats of about 90 amino acids and a C-terminal catalytic domain.</p> <p>The 90 amino-acid repeated domain contains 6 conserved cysteines. It has been shown [1,2] that three disulfide bonds link the first and sixth, second and fifth, and third and fourth cysteines. The domain can be drawn in the shape of an apple (see below) and has been accordingly called the 'apple domain'.</p> <pre> x x x x x x x C---C x x x x x x C x x x x x x x x C x x x x x x x x x x x x x x x x x x x x x x C---C X x..... </pre> <p>Schematic representation of an apple domain.</p> <p>Apart from the cysteines, there are a number of other conserved positions in the apple domain. We have developed a pattern, that spans the complete domain, and which includes these conserved positions.</p>

943

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-x(3)-[LIVMFY]-x(5)-[LIVMFY]-x(3)-[DENQ]-[LIVMFY]-x(10)-C-x(3)-C-T-x(4)-C-x-[LIVMFY]-F-x-[FY]-x(13,14)-C-x-[LIVMFY]-[RK]-x-[ST]-x(14,15)-S-G-x-[ST]-[LIVMFY]-x(2)-C</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update June 1992 / Pattern and text revised.</p> <p>References [1] McMullen B.A., Fujikawa K., Davie E.W. Biochemistry 30:2050-2056(1991).</p> <p>[2] McMullen B.A., Fujikawa K., Davie E.W. Biochemistry 30:2056-2060(1991).</p>
PAP2		PAP2 superfamily	<p>Accession number: PF01569</p> <p>Definition: PAP2 superfamily</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_486 (release 4.0)</p> <p>Gathering cutoffs: 16 16</p> <p>Trusted cutoffs: 22.00 22.00</p> <p>Noise cutoffs: 11.40 11.40</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97194074</p> <p>Reference Title: Identification of a novel phosphatase sequence motif.</p> <p>Reference Author: Stukey J, Carman GM;</p> <p>Reference Location: Protein Sci 1997;6:469-472.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97406916</p> <p>Reference Title: An unexpected structural relationship between integral membrane phosphatases and soluble haloperoxidases.</p> <p>Reference Author: Neuwald AF;</p> <p>Reference Location: Protein Sci 1997;6:1764-1767.</p> <p>Database Reference: INTERPRO; IPR000326;</p> <p>Database reference: PFAMB; PB021113;</p> <p>Database reference: PFAMB; PB040926;</p> <p>Database reference: PFAMB; PB041096;</p> <p>Database reference: PFAMB; PB041301;</p> <p>Comment: This family includes the enzyme type 2 phosphatidic acid phosphatase (PAP2).</p> <p>Number of members: 49</p>
PAPS_reduct		Phosphoadenosine phosphosulfate reductase family	<p>Accession number: PF01507</p> <p>Definition: Phosphoadenosine phosphosulfate reductase family</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_590 (release 4.0)</p> <p>Gathering cutoffs: 49 49</p> <p>Trusted cutoffs: 55.40 55.40</p> <p>Noise cutoffs: -34.60 -34.60</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97411695</p> <p>Reference Title: Crystal structure of phosphoadenylyl sulphate (PAPS) reductase: a new family of adenine nucleotide alpha hydrolases.</p> <p>Reference Title: hydrolases.</p> <p>Reference Author: Savage H, Montoya G, Svensson C, Schwenn JD, Sinning I;</p> <p>Reference Location: Structure 1997;5:895-906.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 96061968</p> <p>Reference Title: Reaction mechanism of thioredoxin:</p> <p>Reference Title: 3'-phospho-adenylylsulfate reductase investigated by site-directed mutagenesis.</p> <p>Reference Title: site-directed mutagenesis.</p> <p>Reference Author: Berendt U, Haverkamp T, Prior A, Schwenn JD;</p>

944

			<p>Reference Location: Eur J Biochem 1995;233:347-356. Reference Number: [3] Reference Medline: 91066949 Reference Title: ATP sulphurylase activity of the nodP and nodQ gene products of Rhizobium meliloti. Reference Author: Schwedock J, Long SR; Reference Location: Nature 1990;348:644-647. Database Reference: SCOP; 1sur; fa; [SCOP-USA][CATH-PDBSUM] Database Reference INTERPRO; IPR002500; Database Reference PDB; 1sur ; 48; 215; Comment: This domain is found in phosphoadenosine phosphosulfate (PAPS) reductase Comment: enzymes or PAPS sulfotransferase. PAPS reductase is part of the adenine Comment: nucleotide alpha hydrolases superfamily also including N type ATP PPases Comment: and ATP sulphurylases [1]. The enzyme uses thioredoxin as an electron Comment: donor for the reduction of PAPS to phospho-adenosine-phosphate (PAP) [1,2]. Comment: It is also found in NodP nodulation protein P from Rhizobium which has ATP Comment: sulphurylase activity (sulfate adenylate transferase) [3]. Number of members: 48</p>
PARP		Poly(ADP-ribose) polymerase catalytic region	<p>Accession number: PF00644 Definition: Poly(ADP-ribose) polymerase catalytic region. Author: Bateman A Alignment method of seed: HMM_built_from_alignment Source of seed members: Bateman A Gathering cutoffs: -59.4 -59.4 Trusted cutoffs: -44.60 -44.60 Noise cutoffs: -180.60 -180.60 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96353841 Reference Title: Structure of the catalytic fragment of poly(AD-ribose) polymerase from chicken. Reference Title: polymerase from chicken. Reference Author: Ruf A, Mennissier de Murcia J, de Murcia G, Schulz GE; Reference Location: Proc Natl Acad Sci U S A 1996;93:7481-7485. Reference Number: [2] Reference Medline: 93293867 Reference Title: The carboxyl-terminal domain of human poly(ADP-ribose) polymerase. Overproduction in Escherichia coli, large scale Reference Title: purification, and characterization. Reference Author: Simonin F, Hofferer L, Panzeter PL, Muller S, de Murcia G, Reference Author: Althaus FR; Reference Location: J Biol Chem 1993;268:13454-13461. Database Reference: SCOP; 1paw; fa; [SCOP-USA][CATH-PDBSUM] Database Reference INTERPRO; IPR001290; Database Reference PDB; 1a26 ; 662; 997; Database Reference PDB; 1pax ; 662; 997; Database Reference PDB; 2pax ; 662; 997; Database Reference PDB; 3pax ; 662; 997; Database Reference PDB; 4pax ; 662; 997; Database Reference PDB; 2paw ; 662; 1009; Database reference: PFAMB; PB041409; Comment: Poly(ADP-ribose) polymerase catalyses the covalent attachment of ADP-ribose units from NAD+ to itself and to a limited number of other DNA binding proteins, which decreases their affinity for DNA. Comment: Poly(ADP-ribose) polymerase is a regulatory component induced by DNA damage. Comment: The carboxyl-terminal region is the most highly conserved region of the protein. Experiments have shown that a carboxyl 40 kDa fragment is still catalytically active [2]. Number of members: 19</p>
PC_rep		Proteasome/cyclosome	<p>Accession number: PF01851 Definition: Proteasome/cyclosome repeat</p>

945

		repeat	<p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 25 0</p> <p>Trusted cutoffs: 30.60 3.00</p> <p>Noise cutoffs: 15.80 15.80</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97348748</p> <p>Reference Title: A repetitive sequence in subunits of the 26S proteasome and</p> <p>Reference Title: 20S cyclosome (anaphase-promoting complex).</p> <p>Reference Author: Lupas A, Baumeister W, Hofmann K;</p> <p>Reference Location: Trends Biochem Sci 1997;22:195-196.</p> <p>Database Reference: INTERPRO; IPR002015;</p> <p>Database reference: PFAMB; PB009978;</p> <p>Database reference: PFAMB; PB040656;</p> <p>Number of members: 112</p>
PE		PE family	<p>Accession number: PF00934</p> <p>Definition: PE family</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_253 (release 3.0)</p> <p>Gathering cutoffs: -20 -20</p> <p>Trusted cutoffs: -10.80 -10.80</p> <p>Noise cutoffs: -20.60 -20.60</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98295987</p> <p>Reference Title: Deciphering the biology of Mycobacterium tuberculosis from</p> <p>Reference Title: the complete genome sequence.</p> <p>Reference Author: Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C,</p> <p>Reference Author: Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd,</p> <p>Reference Author: Tekaia F, Badcock K, Basham D, Brown D,</p> <p>Reference Author: Chillingworth T,</p> <p>Reference Author: Connor R, Davies R, Devlin K, Feltwell T, Gentles S,</p> <p>Reference Author: Hamlin</p> <p>Reference Author: N, Holroyd S, Hornsby T, Jagels K, Barrell BG, et al;</p> <p>Reference Location: Nature 1998;393:537-544.</p> <p>Database Reference: INTERPRO; IPR000084;</p> <p>Comment: This family named after a PE motif near to the amino</p> <p>Comment: terminus of the domain. The PE family of proteins</p> <p>Comment: all contain an amino-terminal region of about 110</p> <p>Comment: amino acids. The carboxyl terminus of this family</p> <p>Comment: are variable and fall into several classes. The</p> <p>Comment: largest class of PE proteins is the highly repetitive</p> <p>Comment: PGRS class which have a high glycine content.</p> <p>Comment: The function of these proteins is uncertain but it</p> <p>Comment: has been suggested that they may be related to</p> <p>Comment: antigenic variation of Mycobacterium tuberculosis [1].</p> <p>Number of members: 90</p>
Pep_deformylase		Polypeptide deformylase	<p>Accession number: PF01327</p> <p>Definition: Polypeptide deformylase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Sarah Teichmann</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 157.40 157.40</p> <p>Noise cutoffs: -29.00 -29.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97002011</p> <p>Reference Title: A new subclass of the zinc metalloproteases superfamily</p> <p>Reference Title: revealed by the solution structure of peptide deformylase.</p> <p>Reference Author: Meinnel T, Blanquet S, Dardel F;</p> <p>Reference Location: J Mol Biol 1996;262:375-386.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98332750</p>

			<p>Reference Title: Solution structure of nickel-peptide deformylase.</p> <p>Reference Author: Dardel F, Ragusa S, Lazennec C, Blanquet S, Meinne T;</p> <p>Reference Location: J Mol Biol 1998;280:501-513.</p> <p>Database Reference: SCOP; 1def; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR000181;</p> <p>Database Reference: PDB; 2def ; 4; 142;</p> <p>Database Reference: PDB; 1def ; 4; 142;</p> <p>Database Reference: PDB; 1dff ; 4; 142;</p> <p>Database Reference: PDB; 1bsj A; 4; 142;</p> <p>Database Reference: PDB; 1bsk A; 4; 142;</p> <p>Database Reference: PDB; 1bs4 A; 4; 142;</p> <p>Database Reference: PDB; 1bs4 B; 504; 642;</p> <p>Database Reference: PDB; 1bs4 C; 1004; 1142;</p> <p>Database Reference: PDB; 1bs5 A; 4; 142;</p> <p>Database Reference: PDB; 1bs5 B; 504; 642;</p> <p>Database Reference: PDB; 1bs5 C; 1004; 1142;</p> <p>Database Reference: PDB; 1bs6 A; 4; 142;</p> <p>Database Reference: PDB; 1bs6 B; 504; 642;</p> <p>Database Reference: PDB; 1bs6 C; 1004; 1142;</p> <p>Database Reference: PDB; 1bs7 A; 4; 142;</p> <p>Database Reference: PDB; 1bs7 B; 504; 642;</p> <p>Database Reference: PDB; 1bs7 C; 1004; 1142;</p> <p>Database Reference: PDB; 1bs8 A; 4; 142;</p> <p>Database Reference: PDB; 1bs8 B; 504; 642;</p> <p>Database Reference: PDB; 1bs8 C; 1004; 1142;</p> <p>Database Reference: PDB; 1bsz A; 4; 142;</p> <p>Database Reference: PDB; 1bsz B; 504; 642;</p> <p>Database Reference: PDB; 1bsz C; 1004; 1142;</p> <p>Database Reference: PDB; 1icj A; 4; 142;</p> <p>Database Reference: PDB; 1icj B; 504; 642;</p> <p>Database Reference: PDB; 1icj C; 1004; 1142;</p> <p>Database reference: PFAMB; PB041251;</p> <p>Number of members: 25</p>
Peptidase_C 15		Pyroglutamyl peptidase	<p>Accession number: PF01470</p> <p>Definition: Pyroglutamyl peptidase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw_manual</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 436.10 436.10</p> <p>Noise cutoffs: -155.40 -155.40</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 99216536</p> <p>Reference Title: The crystal structure of pyroglutamyl peptidase I from</p> <p>Reference Title: bacillus amyloliquefaciens reveals a new structure for a</p> <p>Reference Title: cysteine protease.</p> <p>Reference Author: Odagaki Y, Hayashi A, Okada K, Hirotsu K, Kabashima T, Ito</p> <p>Reference Author: K, Yoshimoto T, Tsuru D, Sato M, Clardy J</p> <p>Reference Location: Structure 1999;7:399-411.</p> <p>Database Reference: SCOP; 1aug; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: MEROPS; C15;</p> <p>Database Reference: INTERPRO; IPR000816;</p> <p>Database Reference: PDB; 1a2z A; 2; 209;</p> <p>Database Reference: PDB; 1a2z B; 2; 209;</p> <p>Database Reference: PDB; 1a2z C; 2; 209;</p> <p>Database Reference: PDB; 1a2z D; 2; 209;</p> <p>Database Reference: PDB; 1aug A; 3; 204;</p> <p>Database Reference: PDB; 1aug B; 213; 414;</p> <p>Database Reference: PDB; 1aug C; 423; 624;</p> <p>Database Reference: PDB; 1aug D; 633; 834;</p> <p>Number of members: 10</p>
Peptidase_M 20	PDOC00613	ArgE / dapE / ACY1 / CPG2 / yscS family signatures	<p>The following enzymes have been shown [1,2,3] to be evolutionary and Functionally related:</p> <p>- In the biosynthetic pathway from glutamate to arginine, the removal of an acetyl group from N2-acetylornithine can be catalyzed via two distinct enzymatic strategies depending on the organism. In some bacteria and in fungi, the acetyl group is transferred on glutamate by glutamate</p>

			<p>acetyltransferase (EC 2.3.1.35) while in enterobacteria such as <i>Escherichia coli</i>, it is hydrolyzed by acetylornithine deacetylase (EC 3.5.1.16) (acetylornithinase) (AO) (gene <i>argE</i>). AO is a homodimeric cobalt-dependent enzyme which displays broad specificity and can also deacylates substrates such as acetylarginine, acetylhistidine, acetylglutamate semialdehyde, etc.</p> <p>- Succinyldiaminopimelate desuccinylase (EC 3.5.1.18) (SDAP) (gene <i>dapE</i>) is the enzyme which catalyzes the fifth step in the biosynthesis of lysine from aspartate semialdehyde: the hydrolysis of succinyl-diaminopimelate to diaminopimelate and succinate. SDAP is an enzyme that requires cobalt or zinc as a cofactor.</p> <p>- Aminoacylase-1 [4] (EC 3.5.1.14) (N-acyl-L-amino-acid amidohydrolase) (ACY1). ACY1 is a homodimeric zinc-binding mammalian enzyme that catalyzes the hydrolysis of N-alpha-acylated amino acids (except for aspartate).</p> <p>- Carboxypeptidase G2 (EC 3.4.17.11) (folate hydrolase G2) (gene <i>cpg2</i>) from <i>Pseudomonas</i> strain RS-16. This enzyme catalyzes the hydrolysis of reduced and non-reduced folates to pterates and glutamate. G2 is a homodimeric zinc-dependent enzyme.</p> <p>- Vacuolar carboxypeptidase S (EC 3.4.17.4) (<i>yscS</i>) from yeast (gene <i>CPS1</i>).</p> <p>- Peptidase T (EC 3.4.11.-) (gene <i>pepT</i>) (tripeptidase) from bacteria. This enzyme catalyzes a variety of tripeptides containing N-terminal methionine, leucine, or phenylalanine.</p> <p>- Xaa-His dipeptidase (EC 3.4.13.3) (carnosinase) from <i>Lactobacillus</i> (gene <i>pepV</i>) [5], a metalloenzyme with activity against beta-alanyl-dipeptides including carnosine (beta-alanyl-histidine).</p> <p>These enzymes share a few characteristics. They hydrolyse peptidic bonds in substrates that share a common structure, they are dependent on cobalt or zinc. For their activity and they are proteins of 40 Kd to 60 Kd with a number of regions of sequence similarity.</p> <p>As signature patterns for these proteins, we selected two of the conserved regions. The first pattern contains a conserved histidine which could be involved in binding metal ions and the second pattern contains a number of conserved charged residues.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIV]-[GALMY]-[LIVMF]-x-[GSA]-H-x-D-[TV]-[STAV] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 6.</p> <p>Consensus pattern [GSTAI]-[SANQ]-D-x-K-[GSACN]-x(2)-[LIVMA]-x(2)-[LIVMFY]-x(14,17)-[LIVM]-x-[LIVMF]-[LIVMSTAG]-[LIVMFA]-x(2)-[DNG]-E-E-x-[GSTN] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note these proteins belong to families M20A/M20B in the classification of peptidases [6,E1]. Last update November 1997 / Patterns and text revised.</p> <p>References</p> <p>[1] Meinzel T., Schmitt E., Mechulam Y., Blanquet S. <i>J. Bacteriol.</i> 174:2323-2331(1992).</p> <p>[2] Boyen A., Charlier D., Sakanyan V., Mett I., Glansdorff N. <i>Gene</i> 116:1-6(1992).</p> <p>[3] Miller C.G., Miller J.L., Bagga D.A. <i>J. Bacteriol.</i> 173:3554-3558(1991).</p> <p>[4] Mitta M., Ohnogi H., Yamamoto A., Kato I., Sakiyama F., Tsunasawa S. <i>J. Biochem.</i> 112:737-742(1992).</p> <p>[5]</p>
--	--	--	---

			<p>Vongerichten K., Klein J., Matern H., Plapp R. Microbiology 140:2591-2600(1994).</p> <p>[6] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>
Peptidase_M 3	PDOC00129	Neutral zinc metallopeptidases, zinc-binding region signature	<p>The majority of zinc-dependent metallopeptidases (with the notable exception Of the carboxypeptidases) share a common pattern of primary structure [1,2,3] in the part of their sequence involved in the binding of zinc, and can be grouped together as a superfamily, known as the metzincins, on the basis of this sequence similarity. They can be classified into a number of distinct families [4,E1] which are listed below along with the proteases which are currently known to belong to these families.</p> <p>Family M1</p> <ul style="list-style-type: none"> - Bacterial aminopeptidase N (EC 3.4.11.2) (gene pepN). - Mammalian aminopeptidase N (EC 3.4.11.2). - Mammalian glutamyl aminopeptidase (EC 3.4.11.7) (aminopeptidase A). It may play a role in regulating growth and differentiation of early B-lineage cells. - Yeast aminopeptidase yscII (gene APE2). - Yeast alanine/arginine aminopeptidase (gene AAP1). - Yeast hypothetical protein YIL137c. - Leukotriene A-4 hydrolase (EC 3.3.2.6). This enzyme is responsible for the hydrolysis of an epoxide moiety of LTA-4 to form LTB-4; it has been shown that it binds zinc and is capable of peptidase activity. <p>Family M2</p> <ul style="list-style-type: none"> - Angiotensin-converting enzyme (EC 3.4.15.1) (dipeptidyl carboxypeptidase I) (ACE) the enzyme responsible for hydrolyzing angiotensin I to angiotensin II. There are two forms of ACE: a testis-specific isozyme and a somatic isozyme which has two active centers. <p>Family M3</p> <ul style="list-style-type: none"> - Thimet oligopeptidase (EC 3.4.24.15), a mammalian enzyme involved in the cytoplasmic degradation of small peptides. - Neurolysin (EC 3.4.24.16) (also known as mitochondrial oligopeptidase M or microsomal endopeptidase). - Mitochondrial intermediate peptidase precursor (EC 3.4.24.59) (MIP). It is involved the second stage of processing of some proteins imported in the mitochondrion. - Yeast saccharolysin (EC 3.4.24.37) (proteinase yscD). - Escherichia coli and related bacteria dipeptidyl carboxypeptidase (EC 3.4.15.5) (gene dcp). - Escherichia coli and related bacteria oligopeptidase A (EC 3.4.24.70) (gene opdA or prIC). - Yeast hypothetical protein YKL134c. <p>Family M4</p> <ul style="list-style-type: none"> - Thermostable thermolysins (EC 3.4.24.27), and related thermolabile neutral proteases (bacillolysins) (EC 3.4.24.28) from various species of Bacillus. - Pseudolysin (EC 3.4.24.26) from Pseudomonas aeruginosa (gene lasB). - Extracellular elastase from Staphylococcus epidermidis. - Extracellular protease prt1 from Erwinia carotovora. - Extracellular minor protease smp from Serratia marcescens. - Vibriolysin (EC 3.4.24.25) from various species of Vibrio. - Protease prtA from Listeria monocytogenes. - Extracellular proteinase proA from Legionella pneumophila. <p>Family M5</p> <ul style="list-style-type: none"> - Mycolysin (EC 3.4.24.31) from Streptomyces cacaui. <p>Family M6</p> <ul style="list-style-type: none"> - Immune inhibitor A from Bacillus thuringiensis (gene ina). Ina degrades two classes of insect antibacterial proteins, attacins and cecropins. <p>Family M7</p> <ul style="list-style-type: none"> - Streptomyces extracellular small neutral proteases

			<p>Family M8</p> <ul style="list-style-type: none"> - Leishmanolysin (EC 3.4.24.36) (surface glycoprotein gp63), a cell surface protease from various species of <i>Leishmania</i>. <p>Family M9</p> <ul style="list-style-type: none"> - Microbial collagenase (EC 3.4.24.3) from <i>Clostridium perfringens</i> and <i>Vibrio alginolyticus</i>.
			<p>Family M10A</p> <ul style="list-style-type: none"> - Serralysin (EC 3.4.24.40), an extracellular metalloprotease from <i>Serratia</i>. - Alkaline metalloproteinase from <i>Pseudomonas aeruginosa</i> (gene aprA). - Secreted proteases A, B, C and G from <i>Erwinia chrysanthemi</i>. - Yeast hypothetical protein YIL108w. <p>Family M10B</p> <ul style="list-style-type: none"> - Mammalian extracellular matrix metalloproteinases (known as matrixins) [5]: MMP-1 (EC 3.4.24.7) (interstitial collagenase), MMP-2 (EC 3.4.24.24) (72 Kd gelatinase), MMP-9 (EC 3.4.24.35) (92 Kd gelatinase), MMP-7 (EC 3.4.24.23) (matrylsin), MMP-8 (EC 3.4.24.34) (neutrophil collagenase), MMP-3 (EC 3.4.24.17) (stromelysin-1), MMP-10 (EC 3.4.24.22) (stromelysin-2), and MMP-11 (stromelysin-3), MMP-12 (EC 3.4.24.65) (macrophage metalloelastase). - Sea urchin hatching enzyme (envelysin) (EC 3.4.24.12). A protease that allows the embryo to digest the protective envelope derived from the egg extracellular matrix. - Soybean metalloendoproteinase 1. <p>Family M11</p> <ul style="list-style-type: none"> - <i>Chlamydomonas reinhardtii</i> gamete lytic enzyme (GLE). <p>Family M12A</p> <ul style="list-style-type: none"> - Astacin (EC 3.4.24.21), a crayfish endoprotease. - Meprin A (EC 3.4.24.18), a mammalian kidney and intestinal brush border metalloendopeptidase. - Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity. The <i>Drosophila</i> homolog of BMP-1 is the dorsal-ventral patterning protein tolloid. - Blastula protease 10 (BP10) from <i>Paracentrotus lividus</i> and the related protein SpAN from <i>Strongylocentrotus purpuratus</i>. - <i>Caenorhabditis elegans</i> protein toh-2. - <i>Caenorhabditis elegans</i> hypothetical protein F42A10.8. - Choriolysins L and H (EC 3.4.24.67) (also known as embryonic hatching proteins LCE and HCE) from the fish <i>Oryzias latipes</i>. These proteases participate in the breakdown of the egg envelope, which is derived from the egg extracellular matrix, at the time of hatching. <p>Family M12B</p> <ul style="list-style-type: none"> - Snake venom metalloproteinases [6]. This subfamily mostly groups proteases that act in hemorrhage. Examples are: adamalysin II (EC 3.4.24.46), atrolysin C/D (EC 3.4.24.42), atrolysin E (EC 3.4.24.44), fibrolase (EC 3.4.24.72), trimerelysin I (EC 3.4.25.52) and II (EC 3.4.25.53). - Mouse cell surface antigen MS2. <p>Family M13</p> <ul style="list-style-type: none"> - Mammalian neprilysin (EC 3.4.24.11) (neutral endopeptidase) (NEP). - Endothelin-converting enzyme 1 (EC 3.4.24.71) (ECE-1), which process the precursor of endothelin to release the active peptide. - Kell blood group glycoprotein, a major antigenic protein of erythrocytes. The Kell protein is very probably a zinc endopeptidase. - Peptidase O from <i>Lactococcus lactis</i> (gene pepO). <p>Family M27</p> <ul style="list-style-type: none"> - Clostridial neurotoxins, including tetanus toxin (TeTx) and the various botulinum toxins (BoNT). These toxins are zinc proteases that block neurotransmitter release by proteolytic cleavage of synaptic proteins such as synaptobrevins, syntaxin and SNAP-25 [7,8]. <p>Family M30</p> <ul style="list-style-type: none"> - <i>Staphylococcus hyicus</i> neutral metalloprotease. <p>Family M32</p>

950

		<p>- Thermolabile carboxypeptidase 1 (EC 3.4.17.19) (carboxypeptidase Taq), an enzyme from <i>Thermus aquaticus</i> which is most active at high temperature.</p> <p>Family M34</p> <p>- Lethal factor (LF) from <i>Bacillus anthracis</i>, one of the three proteins composing the anthrax toxin.</p> <p>Family M35</p> <p>- Deuterolysin (EC 3.4.24.39) from <i>Penicillium citrinum</i> and related proteases from various species of <i>Aspergillus</i>.</p> <p>Family M36</p> <p>- Extracellular elastolytic metalloproteinases from <i>Aspergillus</i>.</p> <p>From the tertiary structure of thermolysin, the position of the residues acting as zinc ligands and those involved in the catalytic activity are known. Two of the zinc ligands are histidines which are very close together in the sequence; C-terminal to the first histidine is a glutamic acid residue which acts as a nucleophile and promotes the attack of a water molecule on the carbonyl carbon of the substrate. A signature pattern which includes the two histidine and the glutamic acid residues is sufficient to detect this superfamily of proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GSTALIVN]-x(2)-H-E-[LIVMFYW]-{DEHRKP}-H-x-[LIVMFYWGSPQ] [The two H's are zinc ligands] [E is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for members of families M5, M7 and M11. Other sequence(s) detected in SWISS-PROT 57; including <i>Neurospora crassa</i> conidiation-specific protein 13 which could be a zinc-protease. Last update July 1999 / Text revised. References [1] Jongeneel C.V., Bouvier J., Bairoch A. FEBS Lett. 242:211-214(1989).</p> <p>[2] Murphy G.J.P., Murphy G., Reynolds J.J. FEBS Lett. 289:4-7(1991).</p> <p>[3] Bode W., Grams F., Reinemer P., Gomis-Rueth F.-X., Baumann U., McKay D.B., Stoecker W. Zoology 99:237-246(1996).</p> <p>[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).</p> <p>[5] Woessner J. Jr. FASEB J. 5:2145-2154(1991).</p> <p>[6] Hite L.A., Fox J.W., Bjarnason J.B. Biol. Chem. Hoppe-Seyler 373:381-385(1992).</p> <p>[7] Montecucco C., Schiavo G. Trends Biochem. Sci. 18:324-327(1993).</p> <p>[8] Niemann H., Blasi J., Jahn R. Trends Cell Biol. 4:179-185(1994).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>	
Peptidase_M 48		Peptidase family M48	Accession number: PF01435 Definition: Peptidase family M48

			<p>Author: Bateman A Alignment method of seed: Clustalw_manual Source of seed members: Swiss-Prot Gathering cutoffs: -35 -35 Trusted cutoffs: -34.00 -34.00 Noise cutoffs: -42.20 -42.20 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Database Reference MEROPS; M48; Database Reference INTERPRO; IPR001915; Database reference: PFAMB; PB008839; Database reference: PFAMB; PB041497; Number of members: 28</p>
Peptidase_S8	PDOC00125	Serine proteases, subtilase family, active sites	<p>Subtilases [1,2] are an extensive family of serine proteases whose catalytic activity is provided by a charge relay system similar to that of the trypsin family of serine proteases but which evolved by independent convergent evolution. The sequence around the residues involved in the catalytic triad (aspartic acid, serine and histidine) are completely different from that of the analogous residues in the trypsin serine proteases and can be used as signatures specific to that category of proteases.</p> <p>The subtilase family currently includes the following proteases:</p> <ul style="list-style-type: none"> - Subtilisins (EC 3.4.21.62), these alkaline proteases from various <i>Bacillus</i> species have been the target of numerous studies in the past thirty years. - Alkaline elastase YaB from <i>Bacillus</i> sp. (gene ale). - Alkaline serine exoprotease A from <i>Vibrio alginolyticus</i> (gene proA). - Aqualysin I from <i>Thermus aquaticus</i> (gene pstI). - AspA from <i>Aeromonas salmonicida</i>. - Bacillopeptidase F (esterase) from <i>Bacillus subtilis</i> (gene bpf). - C5A peptidase from <i>Streptococcus pyogenes</i> (gene scpA). - Cell envelope-located proteases PI, PII, and PIII from <i>Lactococcus lactis</i>. - Extracellular serine protease from <i>Serratia marcescens</i>. - Extracellular protease from <i>Xanthomonas campestris</i>. - Intracellular serine protease (ISP) from various <i>Bacillus</i>. - Minor extracellular serine protease epr from <i>Bacillus subtilis</i> (gene epr). - Minor extracellular serine protease vpr from <i>Bacillus subtilis</i> (gene vpr). - Nisin leader peptide processing protease nisP from <i>Lactococcus lactis</i>. - Serotype-specific antigene 1 from <i>Pasteurella haemolytica</i> (gene ssa1). - Thermitase (EC 3.4.21.66) from <i>Thermoactinomyces vulgaris</i>. - Calcium-dependent protease from <i>Anabaena variabilis</i> (gene prcA). - Halolysin from halophilic bacteria sp. 172p1 (gene hly). - Alkaline extracellular protease (AEP) from <i>Yarrowia lipolytica</i> (gene xpr2). - Alkaline proteinase from <i>Cephalosporium acremonium</i> (gene alp). - Cerevisin (EC 3.4.21.48) (vacuolar protease B) from yeast (gene PRB1). - Cuticle-degrading protease (pr1) from <i>Metarhizium anisopliae</i>. - KEX-1 protease from <i>Kluyveromyces lactis</i>. - Kexin (EC 3.4.21.61) from yeast (gene KEX-2). - Oryzin (EC 3.4.21.63) (alkaline proteinase) from <i>Aspergillus</i> (gene alp). - Proteinase K (EC 3.4.21.64) from <i>Tritirachium album</i> (gene proK). - Proteinase R from <i>Tritirachium album</i> (gene proR). - Proteinase T from <i>Tritirachium album</i> (gene proT). - Subtilisin-like protease III from yeast (gene YSP3). - Thermomycin (EC 3.4.21.65) from <i>Malbranchea sulfurea</i>. - Furin (EC 3.4.21.85), neuroendocrine convertases 1 to 3 (NEC-1 to -3) and PACE4 protease from mammals, other vertebrates, and invertebrates. <p>These proteases are involved in the processing of hormone precursors at sites comprised of pairs of basic amino acid residues [3].</p> <ul style="list-style-type: none"> - Tripeptidyl-peptidase II (EC 3.4.14.10) (tripeptidyl aminopeptidase) from Human. - Prestalk-specific proteins tagB and tagC from slime mold [4]. Both proteins consist of two domains: a N-terminal subtilase catalytic domain and a C-terminal ABC transporter domain (see <PDOC00185>). <p>Description of pattern(s) and/or profile(s)</p>

			<p>Consensus pattern [STAIV]-x-[LIVMF]-[LIVM]-D-[DSTA]-G-[LIVMFC]-x(2,3)-[DNH] [D is the active site residue] Sequences known to belong to this class detected by the pattern the majority of subtilases with a few exceptions. Other sequence(s) detected in SWISS-PROT 44.</p> <p>Consensus pattern H-G-[STM]-x-[VIC]-[STAGC]-[GS]-x-[LIVMA]-[STAGCLV]-[SAGM] [H is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for aspA and ssa1 which both seem to lack the histidine active site. Other sequence(s) detected in SWISS-PROT adenylate cyclase type VIII.</p> <p>Consensus pattern G-T-S-x-[SA]-x-P-x(2)-[STAVC]-[AG] [S is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for nisP, tagC and S.marcescens extracellular serine protease. Other sequence(s) detected in SWISS-PROT 6.</p> <p>Note if a protein includes at least two of the three active site signatures, the probability of it being a serine protease from the subtilase family is 100%</p> <p>Note these proteins belong to family S8 in the classification of peptidases [5,E1]. Expert(s) to contact by email Brannigan J. jab5@vaxa.york.ac.uk</p> <p>Siezen R.J. siezen@nizo.nl</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References [1] Siezen R.J., de Vos W.M., Leunissen J.A.M., Dijkstra B.W. Protein Eng. 4:719-737(1991).</p> <p>[2] Siezen R.J. (In) Proceeding subtilisin symposium, Hamburg, (1992).</p> <p>[3] Barr P.J. Cell 66:1-3(1991).</p> <p>[4] Shaulsky G., Kuspa A., Loomis W.F.; Genes Dev. 9:1111-1122(1995).</p> <p>[5] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>
Peptidase_S9	PDOC00587	Prolyl oligopeptidase family serine active site	<p>The prolyl oligopeptidase family [1,2,3] consist of a number of evolutionary related peptidases whose catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. The known members of this family are listed below.</p> <ul style="list-style-type: none"> - Prolyl endopeptidase (EC 3.4.21.26) (PE) (also called post-proline cleaving enzyme). PE is an enzyme that cleaves peptide bonds on the C-terminal side of prolyl residues. The sequence of PE has been obtained from a mammalian species (pig) and from bacteria (Flavobacterium meningosepticum and Aeromonas hydrophila); there is a high degree of sequence conservation between these sequences. - Escherichia coli protease II (EC 3.4.21.83) (oligopeptidase B) (gene prtB) which cleaves peptide bonds on the C-terminal side of lysyl and arginyl residues. - Dipeptidyl peptidase IV (EC 3.4.14.5) (DPP IV). DPP IV is an enzyme that removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline. - Yeast vacuolar dipeptidyl aminopeptidase A (DPAP A) (gene: STE13) which

			<p>is responsible for the proteolytic maturation of the alpha-factor precursor.</p> <ul style="list-style-type: none"> - Yeast vacuolar dipeptidyl aminopeptidase B (DPAP B) (gene: DAP2). - Acylamino-acid-releasing enzyme (EC 3.4.19.1) (acyl-peptide hydrolase). <p>This enzyme catalyzes the hydrolysis of the amino-terminal peptide bond of an N-acetylated protein to generate a N-acetylated amino acid and a protein with a free amino-terminus.</p> <p>A conserved serine residue has experimentally been shown (in E.coli protease II as well as in pig and bacterial PE) to be necessary for the catalytic mechanism. This serine, which is part of the catalytic triad (Ser, His, Asp), is generally located about 150 residues away from the C-terminal extremity of these enzymes (which are all proteins that contains about 700 to 800 amino acids).</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern D-x(3)-A-x(3)-[LIVMFYW]-x(14)-G-x-S-x-G-G-[LIVMFYW](2) [S is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for yeast DPAP A.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note these proteins belong to families S9A/S9B/S9C in the classification of peptidases [4,E1].</p> <p>Last update November 1997 / Text revised.</p> <p>References</p> <p>[1] Rawlings N.D., Polgar L., Barrett A.J. Biochem. J. 279:907-911(1991).</p> <p>[2] Barrett A.J., Rawlings N.D. Biol. Chem. Hoppe-Seyler 373:353-360(1992).</p> <p>[3] Polgar L., Szabo E. Biol. Chem. Hoppe-Seyler 373:361-366(1992).</p> <p>[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>
Peptidase_U 7		Peptidase family U7	<p>Accession number: PF01343</p> <p>Definition: Peptidase family U7</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_707 (release 2.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 47.60 47.60</p> <p>Noise cutoffs: -55.60 -55.60</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Database Reference MEROPS; U7;</p> <p>Database Reference INTERPRO; IPR002142;</p> <p>Number of members: 37</p>
PEP-utilizers	PDOC00527	PEP-utilizing enzymes signatures	<p>A number of enzymes that catalyze the transfer of a phosphoryl group from phosphoenolpyruvate (PEP) via a phospho-histidine intermediate have been shown to be structurally related [1,2,3,4]. These enzymes are:</p> <ul style="list-style-type: none"> - Pyruvate,orthophosphate dikinase (EC 2.7.9.1) (PPDK). PPDK catalyzes the reversible phosphorylation of pyruvate and phosphate by ATP to PEP and diphosphate. In plants PPDK function in the direction of the formation of PEP, which is the primary acceptor of carbon dioxide in C4 and crassulacean acid metabolism plants. In some bacteria, such as Bacteroides symbiosus,

954

			<p>PPDK functions in the direction of ATP synthesis.</p> <ul style="list-style-type: none"> - Phosphoenolpyruvate synthase (EC 2.7.9.2) (pyruvate, water dikinase). This enzyme catalyzes the reversible phosphorylation of pyruvate by ATP to form PEP, AMP and phosphate, an essential step in gluconeogenesis when pyruvate and lactate are used as a carbon source. - Phosphoenolpyruvate-protein phosphotransferase (EC 2.7.3.9). This is the first enzyme of the phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS), a major carbohydrate transport system in bacteria. The PTS catalyzes the phosphorylation of incoming sugar substrates concomitant with their translocation across the cell membrane. The general mechanism of the PTS is the following: a phosphoryl group from PEP is transferred to enzyme-I (EI) of PTS which in turn transfers it to a phosphoryl carrier protein (HPr). Phospho-HPr then transfers the phosphoryl group to a sugar-specific permease. <p>All these enzymes share the same catalytic mechanism: they bind PEP and transfer the phosphoryl group from it to a histidine residue. The sequence around that residue is highly conserved and can be used as a signature pattern for these enzymes. As a second signature pattern we selected a conserved region in the C-terminal part of the PEP-utilizing enzymes. The biological significance of this region is not yet known.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-[GA]-x-[STN]-x-H-[STA]-[STAV]-[LIVM](2)-[STAV]-[RG] [H is phosphorylated] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [DEQSK]-x-[LIVMF]-S-[LIVMF]-G-[ST]-N-D-[LIVM]-x-Q-[LIVMFYGT]-[STALIV]-[LIVMFY]-[GAS]-x(2)-R Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update December 1999 / Patterns and text revised.</p> <p>References [1] Reizer J., Hoischen C., Reizer A., Pham T.N., Saier M.H. Jr. Protein Sci. 2:506-521(1993).</p> <p>[2] Reizer J., Reizer A., Merrick M.J., Plunkett G. III, Rose D.J., Saier M.H. Jr. Gene 181:103-108(1996).</p> <p>[3] Pocalyko D.J., Carroll L.J., Martin B.M., Babbitt P.C., Dunaway-Mariano D. Biochemistry 29:10757-10765(1990).</p> <p>[4] Niersbach M., Kreuzaler F., Geerse R.H., Postma P., Hirsch H.J. Mol. Gen. Genet. 232:332-336(1992).</p>
PG_binding_2		Putative peptidoglycan binding domain	<p>Accession number: PF01476 Definition: Putative peptidoglycan binding domain Author: Bateman A Alignment method of seed: HMM_built_from_alignment Source of seed members: Bateman A Gathering cutoffs: 22 22 Trusted cutoffs: 22.40 22.10 Noise cutoffs: 21.10 21.10 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 92324582 Reference Title: Modular design of the Enterococcus hirae muramidase-2 and Reference Title: Streptococcus faecalis autolysin. Reference Author: Joris B, Englebert S, Chu CP, Kariyama R, Daneo-Moore L, Reference Author: Shockman GD, Ghuysen JM;</p>

			<p>Reference Location: FEMS Microbiol Lett 1992;70:257-264. Database Reference: INTERPRO; IPR002482; Database reference: PFAMB; PB019287; Database reference: PFAMB; PB040847; Database reference: PFAMB; PB040977; Comment: This domain is about 40 residues long. It is found in a variety Comment: of enzymes involved in bacterial cell wall degradation [1]. This Comment: domain may have a general peptidoglycan binding function. Number of members: 197</p>
phoslip	PDOC00109	Phospholipase A2 active sites signatures	<p>Phospholipase A2 (EC 3.1.1.4) (PA2) [1,2] is an enzyme which releases fatty acids from the second carbon group of glycerol. PA2's are small and rigid proteins of 120 amino-acid residues that have four to seven disulfide bonds. PA2 binds a calcium ion which is required for activity. The side chains of two conserved residues, a histidine and an aspartic acid, participate in a 'catalytic network'.</p> <p>Many PA2's have been sequenced from snakes, lizards, bees and mammals. In the latter, there are at least four forms: pancreatic, membrane-associated as well as two less characterized forms. The venom of most snakes contains multiple forms of PA2. Some of them are presynaptic neurotoxins which inhibit neuromuscular transmission by blocking acetylcholine release from the nerve termini.</p> <p>We derived two different signature patterns for PA2's. The first is centered on the active site histidine and contains three cysteines involved in disulfide bonds. The second is centered on the active site aspartic acid and also contains three cysteines involved in disulfide bonds.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-C-x(2)-H-x(2)-C [H is the active site residue] Sequences known to belong to this class detected by the pattern ALL known functional PA2's. However, this pattern will not detect some snake toxins homologous with PA2 but which have lost their catalytic activity as well as otoconin-22, a Xenopus protein from the aragonitic otoconia which is also unlikely to be enzymatically active. Other sequence(s) detected in SWISS-PROT 15.</p> <p>Consensus pattern [LIVMA]-C-{LIVMFYWPCST}-C-D-x(5)-C [D is the active site residue] Sequences known to belong to this class detected by the pattern the majority of functional and non-functional PA2's. Undetected sequences are bee PA2, gila monster PA2's, PA2 PL-X from habu and PA2 PA-5 from mulga. Other sequence(s) detected in SWISS-PROT 12. Expert(s) to contact by email Seilhamer J.J. jeff@incyte.com</p> <p>Last update November 1995 / Patterns and text revised. References [1] Davidson F.F., Dennis E.A. J. Mol. Evol. 31:228-238(1990). [2] Gomez F., Vandermeers A., Vandermeers-Piret M.-C., Herzog R., Rathe J., Stievenart M., Winand J., Christophe J. Eur. J. Biochem. 186:23-33(1989).</p>
PI3_P14_kinase	PDOC00710	Phosphatidylinositol 3- and 4-kinases signatures	<p>Phosphatidylinositol 3-kinase (PI3-kinase) (EC 2.7.1.137) [1] is an enzyme that phosphorylates phosphoinositides on the 3-hydroxyl group of the inositol ring. The exact function of the three products of PI3-kinase - PI-3-P, PI-3,4-P(2) and PI-3,4,5-P(3) - is not yet known, although it is proposed that they function as second messengers in cell signalling. Currently, three forms of PI3-kinase are known:</p> <p>- The mammalian enzyme which is a heterodimer of a 110 Kd catalytic chain</p>

956

		<p>(p110) and an 85 Kd subunit (p85) which allows it to bind to activated tyrosine protein kinases. There are at least two different types of p100 subunits (alpha and beta).</p> <ul style="list-style-type: none"> - Yeast TOR1/DRR1 and TOR2/DRR2 [2], PI3-kinases required for cell cycle activation. Both are proteins of about 280 Kd. - Yeast VPS34 [3], a PI3-kinase involved in vacuolar sorting and segregation. VPS34 is a protein of about 100 Kd. - Arabidopsis thaliana and soybean VPS34 homologs. <p>Phosphatidylinositol 4-kinase (PI4-kinase) (EC 2.7.1.67) [4] is an enzyme that acts on phosphatidylinositol (PI) in the first committed step in the production of the second messenger inositol-1,4,5,-trisphosphate. Currently the following forms of PI4-kinases are known:</p> <ul style="list-style-type: none"> - Human PI4-kinase alpha. - Yeast PIK1, a nuclear protein of 120 Kd. - Yeast STT4, a protein of 214 Kd. <p>The PI3- and PI4-kinases share a well conserved domain at their C-terminal section; this domain seems to be distantly related to the catalytic domain of protein kinases [2]. We developed two signature patterns from the best conserved parts of this domain.</p> <p>Four additional proteins belong to this family:</p> <ul style="list-style-type: none"> - Mammalian FKBP-rapamycin associated protein (FRAP) [5], which acts as the target for the cell-cycle arrest and immunosuppressive effects of the FKBP12-rapamycin complex. - Yeast protein ESR1 [6] which is required for cell growth, DNA repair and meiotic recombination. - Yeast protein TEL1 which is involved in controlling telomere length. - Yeast hypothetical protein YHR099w, a distantly related member of this family. - Fission yeast hypothetical protein SpAC22E12.16C. <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMFAC]-K-x(1,3)-[DEA]-[DE]-[LIVMC]-R-Q-[DE]-x(4)-Q Sequences known to belong to this class detected by the pattern ALL, except for yeast YHR099w. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [GS]-x-[AV]-x(3)-[LIVM]-x(2)-[FYH]-[LIVM](2)-x-[LIVMF]-x-D-R-H-x(2)-N Sequences known to belong to this class detected by the pattern ALL, except for yeast YHR099w. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References</p> <p>[1] Hiles I.D., Otsu M., Volinia S., Fry M.J., Gout I., Dhand R., Panayotou G., Ruiz-Larrea F., Thompson A., Totty N.F., Hsuan J.J., Courtneidge S.A., Parker P.J., Waterfield M.D. Cell 70:419-429(1992).</p> <p>[2] Kunz J., Henriquez R., Schneider U., Deuter-Reinhard M., Movva N., Hall M.N. Cell 73:585-596(1993).</p> <p>[3] Schu P.V., Takegawa K., Fry M.J., Stack J.H., Waterfield M.D., Emr S.D. Science 260:88-91(1993).</p> <p>[4] Garcia-Bustos J.F., Marini F., Stevenson I., Frei C., Hall M.N. EMBO J. 13:2352-2361(1994).</p> <p>[5] Brown E.J., Albers M.W., Shin T.B., Ichikawa K., Keith C.T., Lane W.S., Schreiber S.L.</p>
--	--	--

957

			<p>Nature 369:756-758(1994).</p> <p>[6] Kato R., Ogawa H. Nucleic Acids Res. 22:3104-3112(1994).</p>
P-II	PDOC00439	P-II protein signatures	<p>The P-II protein (gene glnB) is a bacterial protein important for the control of glutamine synthetase [1,2,3]. In nitrogen-limiting conditions, when the ratio of glutamine to 2-ketoglutarate decreases, P-II is uridylylated on a tyrosine residue to form P-II-UMP. P-II-UMP allows the deadenylation of glutamine synthetase (GS), thus activating the enzyme. Conversely, in nitrogen excess, P-II-UMP is deuridylylated and then promotes the adenylation of GS. P-II also indirectly controls the transcription of the GS gene (glnA) by preventing NR-II (ntrB) to phosphorylate NR-I (ntrC) which is the transcriptional activator of glnA. Once P-II is uridylylated, these events are reversed.</p> <p>P-II is a protein of about 110 amino acid residues extremely well conserved. The tyrosine which is uridylylated is located in the central part of the protein.</p> <p>In cyanobacteria, P-II seems to be phosphorylated on a serine residue rather than being uridylylated.</p> <p>In methanogenic archaeobacteria, the nitrogenase iron protein gene (nifH) is followed by two open reading frames highly similar to the eubacterial P-II protein [4]. These proteins could be involved in the regulation of nitrogen fixation.</p> <p>In the red alga, <i>Porphyra purpurea</i>, there is a glnB homolog encoded in the chloroplast genome.</p> <p>Other proteins highly similar to glnB are:</p> <ul style="list-style-type: none"> - <i>Bacillus subtilis</i> protein nrgB [5]. - <i>Escherichia coli</i> hypothetical protein yba1 [6]. <p>We developed two signature patterns for P-II protein. The first one is a conserved stretch (in eubacteria) of six residues which contains the uridylylated tyrosine, the other is derived from a conserved region in the C-terminal part of the P-II protein.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern Y-[KR]-G-[AS]-[AE]-Y [The second Y is uridylylated] Sequences known to belong to this class detected by the pattern ALL glnB's from eubacteria. Other sequence(s) detected in SWISS-PROT 4.</p> <p>Consensus pattern [ST]-x(3)-G-[DY]-G-[KR]-[IV]-[FW]-[LIVM]-x(2)-[LIVM] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References</p> <p>[1] Magasanik B. Biochimie 71:1005-1012(1989).</p> <p>[2] Hotel A., Merrick M. Mol. Gen. Genet. 215:134-138(1988).</p> <p>[3] Cheah E., Carr P.D., Suffolk P.M., Vasuvedan S.G., Dixon N.E., Ollis D.L. Structure 2:981-990(1994).</p> <p>[4] Sibold L., Henriquet M., Possot O., Aubert J.-P. Res. Microbiol. 142:5-12(1991).</p> <p>[5] Wray L.V. Jr., Atkinson M.R., Fisher S.H. J. Bacteriol. 176:108-114(1994).</p>

[illegible]

959

			<p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Database Reference: SCOP; 1lpa; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database reference: PROSITE_PROFILE; PS50095;</p> <p>Database Reference: INTERPRO; IPR001024;</p> <p>Database Reference: PDB; 1lox ; 2; 112;</p> <p>Database Reference: PDB; 1hpl B; 336; 445;</p> <p>Database Reference: PDB; 1hpl A; 338; 447;</p> <p>Database Reference: PDB; 1eth C; 337; 403;</p> <p>Database Reference: PDB; 1eth A; 339; 405;</p> <p>Database Reference: PDB; 1eth C; 403; 445;</p> <p>Database Reference: PDB; 1eth A; 405; 447;</p> <p>Database Reference: PDB; 1rp1 ; 339; 449;</p> <p>Database Reference: PDB; 1bu8 A; 340; 407;</p> <p>Database Reference: PDB; 1bu8 A; 415; 452;</p> <p>Database Reference: PDB; 1gpl ; 322; 334;</p> <p>Database Reference: PDB; 1ca1 ; 256; 370;</p> <p>Database Reference: PDB; 1qm6 A; 256; 370;</p> <p>Database Reference: PDB; 1qm6 B; 256; 370;</p> <p>Database Reference: PDB; 1qmd A; 256; 370;</p> <p>Database Reference: PDB; 1qmd B; 256; 370;</p> <p>Comment: This domain is found in a variety of membrane or lipid associated proteins. It is called the PLAT</p> <p>Comment: (Polycystin-1, Lipoxxygenase, Alpha-Toxin) domain or LH2 (Lipoxxygenase homology) domain. The known structure</p> <p>Comment: of pancreatic lipase shows this domain binds to</p> <p>Comment: procolipase</p> <p>Comment: Colipase, which mediates membrane association. So it appears possible that this domain mediates</p> <p>Comment: membrane</p> <p>Comment: attachment via other protein binding partners. The</p> <p>Comment: structure of this domain is known for many members of the</p> <p>Comment: family and is composed of a beta sandwich.</p> <p>Number of members: 82</p>
PLRV_ORF5	Potato leaf roll virus readthrough protein	<p>Accession number: PF01690</p> <p>Definition: Potato leaf roll virus readthrough protein</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1335 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 116.40 116.40</p> <p>Noise cutoffs: -285.50 -285.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 94233771</p> <p>Reference Title: Changes in the amino acid sequence of the coat protein readthrough domain of potato leafroll luteovirus affect the</p> <p>Reference Title: formation of an epitope and aphid transmission.</p> <p>Reference Author: Jolly CA, Mayo MA;</p> <p>Reference Location: Virology 1994;201:182-185.</p> <p>Database Reference: INTERPRO; IPR002929;</p> <p>Comment: This family consists mainly of the potato leaf roll virus readthrough protein. This is generated via a readthrough of open reading frame 3 a coat protein allowing</p> <p>Comment: transcription</p> <p>Comment: of open reading frame 5 to give an extended coat protein with a large c-terminal addition or read through domain [1].</p> <p>Comment: The readthrough protein is thought to play a role in the circulative aphid transmission of potato leaf roll virus [1].</p> <p>Comment: Also in the family is open reading frame 6 from beet western</p> <p>Comment: yellows virus and potato leaf roll virus both luteovirus and</p> <p>Comment: an unknown protein from cucurbit aphid-borne yellows virus a</p> <p>Comment: closterovirus.</p> <p>Number of members: 28</p>	
PMSR	Peptide methionine sulfoxide	<p>Accession number: PF01625</p> <p>Definition: Peptide methionine sulfoxide reductase</p> <p>Author: Bateman A</p>	

960

		reductase	<p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1111 (release 4.1)</p> <p>Gathering cutoffs: -62 -62</p> <p>Trusted cutoffs: -28.00 -28.00</p> <p>Noise cutoffs: -96.70 -96.70</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96353931</p> <p>Reference Title: Peptide methionine sulfoxide reductase contributes to the maintenance of adhesins in three major pathogens.</p> <p>Reference Title: maintenance of adhesins in three major pathogens.</p> <p>Reference Author: Wizemann TM, Moskovitz J, Pearce BJ, Cundell D, Arvidson</p> <p>Reference Author: CG, So M, Weissbach H, Brot N, Masure HR;</p> <p>Reference Location: Proc Natl Acad Sci USA 1996;93:7985-7990.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 96312545</p> <p>Reference Title: Cloning the expression of a mammalian gene involved in the</p> <p>Reference Title: reduction of methionine sulfoxide residues in proteins.</p> <p>Reference Author: Moskovitz J, Weissbach H, Brot N;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1996;93:2095-2099.</p> <p>Database Reference: INTERPRO; IPR002569;</p> <p>Comment: This enzyme repairs damaged proteins. Methionine sulfoxide in proteins</p> <p>Comment: is reduced to methionine.</p> <p>Number of members: 28</p>
Pollen_allerg_2		Ribonuclease (pollen allergen)	<p>Accession number: PF01620</p> <p>Definition: Ribonuclease (pollen allergen)</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1050 (release 4.1)</p> <p>Gathering cutoffs: -3 -3</p> <p>Trusted cutoffs: 23.10 23.10</p> <p>Noise cutoffs: -29.40 -29.40</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95246885</p> <p>Reference Title: Major allergen Phl p Vb in timothy grass is a novel pollen RNase.</p> <p>Reference Title: RNase.</p> <p>Reference Author: Bufe A, Schramm G, Keown MB, Schlaak M, Becker WM;</p> <p>Reference Location: Febs lett 1995;363:6-12.</p> <p>Database Reference: INTERPRO; IPR002914;</p> <p>Database reference: PFAMB; PB037130;</p> <p>Comment: This family contains grass pollen proteins of group V.</p> <p>Comment: Swiss:Q40963 has been shown to possess ribonuclease activity [1].</p> <p>Number of members: 27</p>
POR_N		Pyruvate flavodoxin/ferredoxin oxidoreductase (N terminus)	<p>Accession number: PF01855</p> <p>Definition: Pyruvate flavodoxin/ferredoxin oxidoreductase (N terminus)</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_323 (release 4.2)</p> <p>Gathering cutoffs: -116 -116</p> <p>Trusted cutoffs: -113.60 -113.60</p> <p>Noise cutoffs: -119.50 -119.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96125254</p> <p>Reference Title: Molecular and phylogenetic characterization of pyruvate and</p> <p>Reference Title: 2-ketoisovalerate ferredoxin oxidoreductases from</p> <p>Reference Title: Pyrococcus furiosus and pyruvate ferredoxin oxidoreductase</p> <p>Reference Title: from Thermotoga maritima.</p> <p>Reference Author: Kletzin A, Adams MW;</p> <p>Reference Location: J Bacteriol 1996;178:248-257.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 94022264</p>

			<p>Reference Title: Growth of the cyanobacterium Anabaena on molecular nitrogen: NifJ is required when iron is limited.</p> <p>Reference Author: Bauer CC, Scappino L, Haselkorn R;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1993;90:8812-8816.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 99140300</p> <p>Reference Title: Crystal structures of the key anaerobic enzyme pyruvate:ferredoxin oxidoreductase, free and in complex with pyruvate.</p> <p>Reference Author: Chabriere E, Charon MH, Volbeda A, Pieulle L, Hatchikian</p> <p>Reference Author: EC, Fontecilla-Camps JC;</p> <p>Reference Location: Nat Struct Biol 1999;6:182-190.</p> <p>Database Reference: SCOP; 2pda; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: SCOP; 2pda; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002880;</p> <p>Database Reference: PDB; 1b0p A; 43; 328;</p> <p>Database Reference: PDB; 1b0p B; 43; 328;</p> <p>Database Reference: PDB; 2pda A; 43; 328;</p> <p>Database Reference: PDB; 2pda B; 43; 328;</p> <p>Database reference: PFAMB; PB014847;</p> <p>Comment: This family includes the N terminal region of the pyruvate ferredoxin</p> <p>Comment: oxidoreductase, corresponding to the first two structural domains.</p> <p>Comment: This region is involved in inter subunit contacts [3].</p> <p>Pyruvate</p> <p>Comment: oxidoreductase (POR) catalyses the final step in the fermentation</p> <p>Comment: of carbohydrates in anaerobic microorganisms [1]. This involves the</p> <p>Comment: oxidative decarboxylation of pyruvate with the participation of</p> <p>Comment: thiamine followed by the transfer of an acetyl moiety to coenzyme</p> <p>Comment: A for the synthesis of acetyl-CoA [1]. The family also includes</p> <p>Comment: pyruvate flavodoxin oxidoreductase as encoded by the nifJ gene in</p> <p>Comment: cyanobacterium which is required for growth on molecular nitrogen</p> <p>Comment: when iron is limited [2].</p> <p>Number of members: 55</p>
PPE		PPE family	<p>Accession number: PF00823</p> <p>Definition: PPE family</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw_manual</p> <p>Source of seed members: Pfam-B_297 (release 3.0)</p> <p>Gathering cutoffs: -90 -90</p> <p>Trusted cutoffs: -88.20 -88.20</p> <p>Noise cutoffs: -105.30 -105.30</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98295987</p> <p>Reference Title: Deciphering the biology of Mycobacterium tuberculosis from</p> <p>Reference Title: the complete genome sequence.</p> <p>Reference Author:</p> <p>Reference Location: Nature 1998;393:537-544.</p> <p>Database Reference: INTERPRO; IPR000030;</p> <p>Database reference: PFAMB; PB040834;</p> <p>Comment: This family named after a PPE motif near to the amino terminus of the domain. The PPE family of proteins all contain an amino-terminal region of about 180</p> <p>Comment: amino acids. The carboxyl terminus of this family are variable, and on the basis of this region fall</p> <p>Comment: into at least three groups. The MPTR subgroup has tandem copies of a motif NXGXGNXG. The second</p> <p>subgroup</p> <p>Comment: contains a conserved motif at about position 350.</p> <p>Comment: The third group are only related in the amino terminal region.</p> <p>Comment:</p>

			<p>Comment: The function of these proteins is uncertain but it has been suggested that they may be related to antigenic variation of Mycobacterium tuberculosis [1].</p> <p>Number of members: 75</p>
PRA-CH		Phosphoribosyl-AMP cyclohydrolase	<p>Accession number: PF01502</p> <p>Definition: Phosphoribosyl-AMP cyclohydrolase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_782 (release 4.0)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 88.20 88.20</p> <p>Noise cutoffs: -44.30 -44.30</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 99129952</p> <p>Reference Title: N1-(5'-phosphoribosyl)adenosine-5'-monophosphate cyclohydrolase: purification and characterization of a unique metalloenzyme.</p> <p>Reference Author: D'Ordine RL, Klem TJ, Davisson VJ;</p> <p>Reference Location: Biochemistry 1999;38:1537-1546.</p> <p>Database Reference: INTERPRO; IPR002496;</p> <p>Comment: This enzyme catalyses the third step in the histidine biosynthetic pathway. It requires Zn ions for activity.</p> <p>Number of members: 28</p>
PRA-PH		Phosphoribosyl-ATP pyrophosphohydrolase	<p>Accession number: PF01503</p> <p>Definition: Phosphoribosyl-ATP pyrophosphohydrolase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_784 (release 4.0)</p> <p>Gathering cutoffs: 6 6</p> <p>Trusted cutoffs: 12.10 12.10</p> <p>Noise cutoffs: 1.00 1.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 79216449</p> <p>Reference Title: The product of the his4 gene cluster in Saccharomyces cerevisiae. A trifunctional polypeptide.</p> <p>Reference Author: Keesey JK Jr, Bigelis R, Fink GR;</p> <p>Reference Location: J Biol Chem 1979 Aug 10;254:7427-7433.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 86310274</p> <p>Reference Title: Primary and secondary structural homologies between the HIS4 gene product of Saccharomyces cerevisiae and the hisE and hisD gene products of Escherichia coli and Salmonella typhimurium.</p> <p>Reference Author: Bruni CB, Carlomagno MS, Formisano S, Paoletta G;</p> <p>Reference Location: Mol Gen Genet 1986;203:389-396.</p> <p>Database Reference: INTERPRO; IPR002497;</p> <p>Comment: This enzyme catalyses the second step in the histidine biosynthetic pathway.</p> <p>Number of members: 32</p>
PseudoU_synth_1		tRNA pseudouridine synthase	<p>Accession number: PF01416</p> <p>Definition: tRNA pseudouridine synthase</p> <p>Previous Pfam IDs: PseudoU_synt;</p> <p>Author: Howe K</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: swissprot</p> <p>Gathering cutoffs: 30 30</p> <p>Trusted cutoffs: 39.10 39.10</p> <p>Noise cutoffs: -55.00 -55.00</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98254513</p> <p>Reference Title: Transfer RNA-pseudouridine synthetase Pus1 of Saccaromyces</p>

			<p>Reference Title: cerevisiae contains one atom of zinc essential for its native conformation and tRNA recognition.</p> <p>Reference Author: Arluison V, Hountondji C, Robert B, Grosjean H;</p> <p>Reference Location: Biochemistry 1998;37:7268-7276.</p> <p>Database Reference: INTERPRO; IPR001406;</p> <p>Database reference: PFAMB; PB027500;</p> <p>Comment: Involved in the formation of pseudouridine at the anticodon stem</p> <p>Comment: and loop of transfer-RNAs</p> <p>Comment: Pseudouridine is an isomer of uridine (5-(beta-D-ribofuranosyl)</p> <p>Comment: uracil, and is the most abundant modified nucleoside found in</p> <p>Comment: all cellular RNAs.</p> <p>Comment: The TruA-like proteins also exhibit a conserved sequence with</p> <p>Comment: a strictly conserved aspartic acid, likely involved in catalysis</p> <p>Number of members: 31</p>
PseudoU_syn th_2		RNA pseudouridylate synthase	<p>Accession number: PF00849</p> <p>Definition: RNA pseudouridylate synthase</p> <p>Previous Pfam IDs: YABO;</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_421 (release 3.0)</p> <p>Gathering cutoffs: 20 20</p> <p>Trusted cutoffs: 20.90 20.90</p> <p>Noise cutoffs: -44.40 -44.40</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96079974</p> <p>Reference Title: A dual-specificity pseudouridine synthase: an Escherichia coli synthase purified and cloned on the basis of its specificity for psi 746 in 23S RNA is also specific for psi 32 in tRNA(phe).</p> <p>Reference Author: Wrzesinski J, Nurse K, Bakin A, Lane BG, Ofengand J;</p> <p>Reference Location: RNA 1995;1:437-448.</p> <p>Database Reference: PROSITE; PDOC00869</p> <p>Database Reference: PROSITE; PDOC00885</p> <p>Database Reference: INTERPRO; IPR000613;</p> <p>Database reference: PFAMB; PB041160;</p> <p>Database reference: PFAMB; PB041232;</p> <p>Comment: Members of this family are involved in modifying bases in RNA molecules.</p> <p>Comment: They carry out the conversion of uracil bases to pseudouridine. This family</p> <p>Comment: includes RluD Swiss:P33643, a pseudouridylate synthase that converts</p> <p>Comment: specific uracils to pseudouridine in 23S rRNA. RluA from E. coli</p> <p>Comment: converts bases in both rRNA and tRNA [1].</p> <p>Number of members: 78</p>
PWI		PWI domain	<p>Accession number: PF01480</p> <p>Definition: PWI domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw_manual</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 64.40 64.40</p> <p>Noise cutoffs: -3.50 -3.50</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 10322432</p> <p>Reference Title: The PWI motif: a new protein domain in splicing factors.</p> <p>Reference Author: Blencowe BJ, Ouzounis CA;</p> <p>Reference Location: Trends Biochem Sci 1999;24:179-180.</p> <p>Database Reference: INTERPRO; IPR002483;</p> <p>Number of members: 11</p>
R3H		R3H domain	<p>Accession number: PF01424</p>

			<p>Definition: R3H domain Author: Bateman A Alignment method of seed: Manual Source of seed members: Medline:99003905 Gathering cutoffs: 25 25 Trusted cutoffs: 59.30 59.30 Noise cutoffs: 5.10 5.10 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 99003905 Reference Title: The R3H motif: a domain that binds single-stranded nucleic acids. Reference Author: Grishin NV; Reference Location: Trends Biochem Sci 1998;23:329-330. Database Reference: INTERPRO: IPR001374; Database reference: PFAM; PF041444; Comment: The name of the R3H domain comes from the characteristic spacing Comment: of the most conserved arginine and histidine residues. The Comment: function of the domain is predicted to be binding ssDNA. Number of members: 28</p>
RepB_protein		Initiator RepB protein	<p>Accession number: PF01051 Definition: Initiator RepB protein Author: Finn RD, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_313 (release 3.0) Gathering cutoffs: 14 14 Trusted cutoffs: 19.00 16.20 Noise cutoffs: 11.80 12.90 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98284148 Reference Title: Replication and control of circular bacterial plasmids. Reference Author: del Solar G, Giraldo R, Ruiz-Echevarria MJ, Espinosa M, Reference Author: Diaz-Orejas R; Reference Location: Microbiol Mol Biol Rev 1998;62:434-464. Reference Number: [2] Reference Medline: 97324207 Reference Title: Initiation of replication of plasmid pMV158: mechanisms of Reference Title: DNA strand-transfer reactions mediated by the initiator RepB protein. Reference Author: Moscoso M, Eritja R, Espinosa M; Reference Location: J Mol Biol 1997;268:840-856. Database Reference: INTERPRO: IPR000525; Database Reference: PDB; 1rep C; 198; 240; Database reference: PFAM; PF000509; Comment: This protein is an initiator of plasmid replication. Comment: RepB possesses nicking-closing (topoisomerase I) like activity. Comment: It is also able to perform a strand transfer reaction on ssDNA Comment: that contains its target. Number of members: 51</p>
Rhomboid		Rhomboid family	<p>Accession number: PF01694 Definition: Rhomboid family Author: Sohrmann M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1399 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 143.60 143.60 Noise cutoffs: -43.60 -43.60 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 90249726 Reference Title: rhomboid, a gene required for dorsoventral axis</p>

965

			<p>Reference Title: establishment and peripheral nervous system development in</p> <p>Reference Title: Drosophila melanogaster.</p> <p>Reference Author: Bier E, Jan LY, Jan YN;</p> <p>Reference Location: Genes Dev 1990;4:190-203.</p> <p>Database Reference INTERPRO; IPR002610;</p> <p>Database reference: PFAMB; PB041113;</p> <p>Comment: This family contains integral membrane proteins that are related to Drosophila rhomboid protein Swiss:P20350.</p> <p>Comment: Members</p> <p>Comment: of this family are found in bacteria and eukaryotes. These proteins contain three strongly conserved histidines in the putative transmembrane regions that may be involved in the</p> <p>Comment: as yet unknown function of these proteins.</p> <p>Number of members: 27</p>
Ribosomal_L 18ae		Ribosomal L18ae protein family	<p>Accession number: PF01775</p> <p>Definition: Ribosomal L18ae protein family</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: PSI-BLAST Q02543</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 136.70 136.70</p> <p>Noise cutoffs: -99.80 -99.80</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Database Reference INTERPRO; IPR002670;</p> <p>Number of members: 11</p>
Ribosomal_L 21p	PDOC00899	Ribosomal protein L21 signature	<p>Ribosomal protein L21 is one of the proteins from the large ribosomal subunit. In Escherichia coli, L21 is known to bind to the 23S rRNA in the presence of L20. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities, groups:</p> <ul style="list-style-type: none"> - Eubacterial L21. - Marchantia polymorpha chloroplast L21. - Cyanelle L21. - Spinach chloroplast L21 (nuclear-encoded). <p>Eubacterial L21 is a protein of about 100 amino-acid residues, the mature form of the spinach chloroplast L21 has 200 residues. As a signature pattern, we selected a conserved region located in the C-terminal section of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [IVT]-x(3)-[KR]-x(3)-[KRQ]-K-x(6)-G-[HF]-R-[RQ]-x(2)-[ST]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update</p> <p>July 1999 / Pattern and text revised.</p>
Ribosomal_L 22e		Ribosomal L22e protein family	<p>Accession number: PF01776</p> <p>Definition: Ribosomal L22e protein family</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: PSI-BLAST P56628</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 262.80 262.80</p> <p>Noise cutoffs: -52.00 -52.00</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Database Reference INTERPRO; IPR002671;</p> <p>Number of members: 11</p>
Ribosomal_L 27e		Ribosomal L27e protein family	<p>Accession number: PF01777</p> <p>Definition: Ribosomal L27e protein family</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: PSI-BLAST P51419</p> <p>Gathering cutoffs: 25 25</p>

966

			<p>Trusted cutoffs: 326.90 326.90 Noise cutoffs: -47.80 -47.80 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Database Reference INTERPRO; IPR001141; Number of members: 9</p>
Ribosomal_L 29	PDOC00501	Ribosomal protein L29 signature	<p>Ribosomal protein L29 is one of the proteins from the large ribosomal subunit. L29 belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:</p> <ul style="list-style-type: none"> - Eubacterial L29. - Red algal L29. - Archaeobacterial L29. - Mammalian L35 - Caenorhabditis elegans L35 (ZK652.4). - Yeast L35. <p>L29 is a protein of 63 to 138 amino-acid residues. As a signature pattern, we selected a conserved region located in the central section of L29.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KNQS]-[PSTLN]-x(2)-[LIMFA]-[KRGSA]-x-[LIVYSTA]-[KR]-[KRHQS]-[DESTANRL]-[LIV]-A-[KRCQVT]-[LIVMA] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 2. Last update December 1999 / Pattern and text revised. References [1] Otaka E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).</p>
Ribosomal_L 31e	PDOC00881	Ribosomal protein L31e signature	<p>A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:</p> <ul style="list-style-type: none"> - Mammalian L31 [1]. - Chlamydomonas reinhardtii L31. - Yeast L34. - Halobacterium marismortui HL30 [2]. <p>These proteins have 87 to 128 amino-acid residues. As a signature pattern, we selected a conserved region located in the central section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern V-[KR]-[LIVM]-x(3)-[LIVM]-N-x-[AKH]-x-W-x-[KR]-G Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update July 1999 / Pattern and text revised. References [1] Tanaka T., Kuwano Y., Kuzumaki T., Ishikawa K., Ogata K. Eur. J. Biochem. 162:45-48(1987).</p> <p>[2] Bergmann U., Arndt E. Biochim. Biophys. Acta 1050:56-60(1990).</p>
Ribosomal_L 35Ae	PDOC00849	Ribosomal protein L35Ae signature	<p>A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:</p> <ul style="list-style-type: none"> - Vertebrate L35A. - Caenorhabditis elegans L35A (F10E7.7). - Yeast L37A/L37B (Rp47). - Pyrococcus woesei L35A homolog [1].

			<p>These proteins have 87 to 110 amino-acid residues. As a signature pattern, we selected a highly conserved stretch of 22 residues in the C-terminal part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-K-[LIVM]-x-R-x-H-G-x(2)-G-x-V-x-A-x-F-x(3)-[LI]-P Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update November 1997 / Pattern and text revised. References [1] Ouzounis C., Kyripides N., Sander C. Nucleic Acids Res. 23:565-570(1995).</p>
Ribosomal_L 35p	PDOC00721	Ribosomal protein L35 signature	<p>Ribosomal protein L35 is one of the proteins from the large subunit of the ribosome. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:</p> <ul style="list-style-type: none"> - Eubacterial L35. - Plant chloroplast L35 (nuclear-encoded). - Red algal chloroplast L35. - Cyanelle L35. <p>L35 is a basic protein of 60 to 70 amino-acid residues. As a signature pattern we selected a conserved region in the N-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-K-[TV]-x(2)-[GSA]-[SAILV]-x-K-R-[LIVMFY]-[KRLS] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update December 1999 / Pattern and text revised. References [1] Otaka E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).</p>
Ribosomal_L 36e	PDOC00916	Ribosomal protein L36e signature	<p>A number of eukaryotic ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:</p> <ul style="list-style-type: none"> - Mammalian L36 [1]. - Drosophila L36 (M(1)1B). - Caenorhabditis elegans L36 (F37C12.4). - Candida albicans L39. - Yeast YL39. <p>These proteins have 99 to 104 amino acids. As a signature pattern, we selected a conserved region in the central part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-Y-E-[KR]-R-x-[LIVM]-[DE]-[LIVM](2)-[KR] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update November 1997 / First entry. References [1] Chan Y.-L., Paz V., Olvera J., Wool I.G. Biochem. Biophys. Res. Commun. 192:849-853(1993).</p>
Ribosomal_L 37ae		Ribosomal L37ae protein family	<p>Accession number: PF01780 Definition: Ribosomal L37ae protein family Author: Bateman A</p>

CONFIDENTIAL

			<p>Alignment method of seed: Clustalw Source of seed members: PSI-BLAST P54051 Gathering cutoffs: 25 25 Trusted cutoffs: 145.10 145.10 Noise cutoffs: -46.90 -46.90 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Database Reference INTERPRO; IPR002674; Comment: This ribosomal protein is found in archaeobacteria and Comment: eukaryotes. It contains four conserved cysteine Comment: residues that may bind to zinc. Number of members: 15</p>
Ribosomal_L 37e	PDOC00827	Ribosomal protein L37e signature	<p>A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:</p> <ul style="list-style-type: none"> - Mammalian L37 [1]. - Leishmania infantum L37 [2]. - Fission yeast YL35 [3]. - Halobacterium marismortui L37e (L35e) [4]. <p>These proteins have 56 to 96 amino-acid residues. As a signature pattern, we selected a highly conserved region located in the N-terminal part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-T-x-[SA]-x-G-x-[KR]-x(3)-[STLR]-x(0,1)-H-x(2)-C-x-R-C-G Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update July 1999 / Pattern and text revised. References [1] Chan Y.-L., Paz V., Olvera J., Wool I.G. Biochem. Biophys. Res. Commun. 192:590-596(1993).</p> <p>[2] Myler P.J., Tripp C.A., Thomas L., Venkataraman G.M., Merlin G., Stuart K. Mol. Biochem. Parasitol. 62:147-152(1993).</p> <p>[3] Otaka E., Higo K.-I., Itoh T. Mol. Gen. Genet. 191:519-524(1983).</p> <p>[4] Bergmann U., Wittmann-Liebold B. Biochim. Biophys. Acta 1173:195-200(1993).</p>
Ribosomal_L 38e		Ribosomal L38e protein family	<p>Accession number: PF01781 Definition: Ribosomal L38e protein family Author: Bateman A Alignment method of seed: Clustalw Source of seed members: PSI-BLAST P23411 Gathering cutoffs: 25 25 Trusted cutoffs: 127.60 127.60 Noise cutoffs: -24.50 -24.50 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 91207349 Reference Title: The primary structure of rat ribosomal protein L38. Reference Author: Kuwano Y, Olvera J, Wool IG; Reference Location: Biochem Biophys Res Commun 1991;175:551-555. Database Reference INTERPRO; IPR002675; Number of members: 8</p>
Ribosomal_L 39	PDOC00050	Ribosomal protein L39e signature	<p>A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:</p>

969

			<ul style="list-style-type: none"> - Mammalian L39 [1]. - Plants L39. - Yeast L46 [2]. - Archebacterial L39e [3]. <p>These proteins are very basic. About 50 residues long, they are the smallest proteins of eukaryotic-type ribosomes. As a signature pattern, we selected a conserved region in the C-terminal section of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KRA]-T-x(3)-[LIVM]-[KRQF]-x-[NHS]-x(3)-R-[NHY]-W-R-R</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1998 / Pattern and text revised.</p> <p>References [1] Lin A., McNally J., Wool I.G. J. Biol. Chem. 259:487-490(1984).</p> <p>[2] Leer R.J., van Raamsdonk-Duin M.M.C., Kraakman P., Mager W.H., Planta R.J. Nucleic Acids Res. 13:701-709(1985).</p> <p>[3] Ramirez C., Louie K.A., Matheson A.T. FEBS Lett. 250:416-418(1989).</p>
Ribosomal_L 4	PDOC00724	Ribosomal protein L1e signature	<p>A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists [1,2,3,4] of:</p> <ul style="list-style-type: none"> - Vertebrate L1 (L4). - Drosophila L1. - Plant L1. - Yeast L2 (Rp2). - Fission yeast L2. - Halobacterium marismortui HmaL4 (HL6). - Methanococcus jannaschii MJ0177. <p>These proteins have 246 (archaeobacteria) to 427 (human) amino acids. As a signature pattern, we selected a conserved region in the N-terminal part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern N-x(3)-[KRM]-x(2)-A-[LIVT]-x-S-A-[LIV]-x-A-[ST]-[SGA]-x(7)-[RK]-[GS]-H</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Rafti F., Gargiulo G., Manzi A., Malva C., Graziani F. Nucleic Acids Res. 17:456-456(1989).</p> <p>[2] Presutti C., Villa T., Bozzoni I. Nucleic Acids Res. 21:3900-3900(1993).</p> <p>[3] Bagni C., Mariottini P., Annesi F., Amaldi F., Arndt E., Kroemer W., Hatakeyama T. Biochim. Biophys. Acta 1216:475-478(1993). J. Biol. Chem. 265:3034-3039(1990).</p>
Ribosomal_S		Ribosomal	Accession number: PF01649

970

20p		protein S20	<p>Definition: Ribosomal protein S20</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1685 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 57.30 57.30</p> <p>Noise cutoffs: -25.50 -25.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 88230452</p> <p>Reference Title: Interaction of proteins S16, S17 and S20 with 16 S ribosomal RNA.</p> <p>Reference Author: Stern S, Changchien LM, Craven GR, Noller HF;</p> <p>Reference Location: J Mol Biol 1988;200:291-299.</p> <p>Database Reference: INTERPRO; IPR002583;</p> <p>Comment: Bacterial ribosomal protein S20 interacts with 16S rRNA [1].</p> <p>Number of members: 29</p>
Ribosomal_S 27e	PDOC00898	Ribosomal protein S27e signature	<p>A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of [1]:</p> <ul style="list-style-type: none"> - Mammalian S27 (human S27 was originally known as metallopan-stimulin 1). - Chlamydomonas reinhardtii S27. - Entamoeba histolytica S27. - Yeast S27. - Archaeobacterial S27e. <p>These proteins have from 62 to 87 amino acids. They contain, in their central section, a putative zinc-finger region of the type C-x(2)-C-x(14)-C-x(2)-C. We have selected that region as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [QKT]-C-x(2)-C-x(6)-F-[GSD]-x-[PSA]-x(5)-C-x(2)-C-[GSA]-x(2)-[LV]-x(2)-P-x-G [The four C's are potential zinc ligands]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update</p> <p>December 1999 / Pattern and text revised.</p> <p>References</p> <p>[1]</p> <p>Chan Y.-L., Suzuki K., Olvera J., Wool I.G.</p> <p>Nucleic Acids Res. 21:649-655(1993).</p>
Ribosomal_S 3_C	PDOC00474	Ribosomal protein S3 signature	<p>Ribosomal protein S3 is one of the proteins from the small ribosomal subunit. In Escherichia coli, S3 is known to be involved in the binding of initiator Met-tRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:</p> <ul style="list-style-type: none"> - Eubacterial S3. - Algal and plant chloroplast S3. - Cyanelle S3. - Archaeobacterial S3. - Plant mitochondrial S3. - Vertebrate S3. - Insect S3. - Caenorhabditis elegans S3 (C23G10.3). - Yeast S3 (Rp13). <p>S3 is a protein of 209 to 559 amino-acid residues. As signature patterns, we selected a conserved region located in the C-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GSTA]-[KR]-x(6)-G-x-[LIVMT]-x(2)-[NQSCH]-x(1,3)-[LIVFCA]-x(3)-[LIV]-[DENQ]-x(7)-[LMT]-x(2)-G-x(2)-[GS]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for</p>

971

			<p>some mitochondrial S3. Other sequence(s) detected in SWISS-PROT NONE. Expert(s) to contact by email Hallick R.B. hallick@arizona.edu</p> <p>Last update December 1999 / Pattern and text revised. References [1] Otaka E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).</p>
Ribosomal_S 3_N	PDOC00474	Ribosomal protein S3 signature	<p>Ribosomal protein S3 is one of the proteins from the small ribosomal subunit. In Escherichia coli, S3 is known to be involved in the binding of initiator Met-tRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:</p> <ul style="list-style-type: none"> - Eubacterial S3. - Algal and plant chloroplast S3. - Cyanelle S3. - Archaeobacterial S3. - Plant mitochondrial S3. - Vertebrate S3. - Insect S3. - Caenorhabditis elegans S3 (C23G10.3). - Yeast S3 (Rp13). <p>S3 is a protein of 209 to 559 amino-acid residues. As signature patterns, we selected a conserved region located in the C-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GSTA]-[KR]-x(6)-G-x-[LIVMT]-x(2)-[NQSCH]-x(1,3)-[LIVFCA]-x(3)-[LIV]-[DENQ]-x(7)-[LMT]-x(2)-G-x(2)-[GS] Sequences known to belong to this class detected by the pattern ALL, except for some mitochondrial S3. Other sequence(s) detected in SWISS-PROT NONE. Expert(s) to contact by email Hallick R.B. hallick@arizona.edu</p> <p>Last update December 1999 / Pattern and text revised. References [1] Otaka E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).</p>
RimM		RimM	<p>Accession number: PF01782 Definition: RimM Author: Bateman A Alignment method of seed: Clustalw Source of seed members: PSI-BLAST P51419 Gathering cutoffs: 25 25 Trusted cutoffs: 49.00 49.00 Noise cutoffs: -66.10 -66.10 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98083058 Reference Title: RimM and RbfA are essential for efficient processing of 16S Reference Title: rRNA in Escherichia coli. Reference Author: Bylund GO, Wipemo LC, Lundberg LA, Wikstrom PM; Reference Location: J Bacteriol 1998;180:73-82. Database Reference INTERPRO; IPR002676; Comment: The RimM protein is essential for efficient processing of 16S rRNA [1]. Comment: The RimM protein was shown to have affinity for free ribosomal 30S Comment: subunits but not for 30S subunits in the 70S ribosomes [1]. Number of members: 14</p>

RNA_dep_R NA_pol		RNA dependent RNA polymerase	<p>Accession number: PF00680</p> <p>Definition: RNA dependent RNA polymerase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_32 (release 2.1)</p> <p>Gathering cutoffs: -127 -127</p> <p>Trusted cutoffs: -117.00 -117.00</p> <p>Noise cutoffs: -137.30 -137.30</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Database Reference: SCOP; 1rdr; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR001205;</p> <p>Database Reference: PDB; 1rdr ; 12; 37;</p> <p>Database Reference: PDB; 1rdr ; 182; 460;</p> <p>Database Reference: PDB; 1rdr ; 67; 97;</p> <p>Database reference: PFAMB; PB039844;</p> <p>Database reference: PFAMB; PB040630;</p> <p>Database reference: PFAMB; PB040631;</p> <p>Database reference: PFAMB; PB040844;</p> <p>Database reference: PFAMB; PB041022;</p> <p>Database reference: PFAMB; PB041498;</p> <p>Number of members: 271</p>
RNA_dep_RN Apol2		RNA dependent RNA polymerase	<p>Accession number: PF00978</p> <p>Definition: RNA dependent RNA polymerase</p> <p>Author: Finn RD, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_13 (release 3.0)</p> <p>Gathering cutoffs: 8.5 0</p> <p>Trusted cutoffs: 8.50 0.20</p> <p>Noise cutoffs: 8.40 8.40</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 93188140</p> <p>Reference Title: Roles of nonstructural polyproteins and cleavage products in regulating Sindbis virus RNA replication and transcription.</p> <p>Reference Author: Lemm JA, Rice CM;</p> <p>Reference Location: J Virol 1993;67:1916-1926.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 96323143</p> <p>Reference Title: Complete replication in vitro of tobacco mosaic virus RNA by a template-dependent, membrane-bound RNA polymerase.</p> <p>Reference Author: Osman TA, Buck KW;</p> <p>Reference Location: J Virol 1996;70:6227-6234.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 94047331</p> <p>Reference Title: Bromovirus RNA replication and transcription require compatibility between the polymerase- and helicase-like viral RNA synthesis proteins.</p> <p>Reference Author: Dinant S, Janda M, Kroner PA, Ahlquist P;</p> <p>Reference Location: J Virol 1993;67:7181-7189.</p> <p>Reference Number: [4]</p> <p>Reference Medline: 94094568</p> <p>Reference Title: Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences.</p> <p>Reference Author: Koonin EV, Dolja VV;</p> <p>Reference Location: Crit Rev Biochem Mol Biol 1993;28:375-430.</p> <p>Database Reference: INTERPRO; IPR001788;</p> <p>Database reference: PFAMB; PB000096;</p> <p>Database reference: PFAMB; PB006751;</p> <p>Comment: This family may represent an RNA dependent RNA polymerase.</p> <p>Comment: The family contains the following proteins:</p> <p>Comment: 2A protein from bromoviruses</p> <p>Comment: putative RNA dependent RNA polymerase from tobamoviruses</p> <p>Comment: Non structural polyprotein from togaviruses</p> <p>Number of members: 125</p>

973

RNA_pol	PDOC00410	Bacteriophage-type RNA polymerase family active site signatures	<p>Many forms of RNA polymerase (EC 2.7.7.6) are known. Most RNA polymerases are multimeric enzymes, but there is a family of single chain polymerases, which are evolutionary related, and which originate from bacteriophages or from mitochondria. The RNA polymerases that belong to this family are [1]:</p> <ul style="list-style-type: none"> - Podoviridae bacteriophages T3, T7, and K11 polymerase. - Bacteriophage SP6 polymerase. - Vertebrate mitochondrial polymerase (gene POLRMT). - Fungal mitochondrial polymerase (gene RPO41). - Polymerases encoded on mitochondrial linear DNA plasmids in various fungi and plants: <i>Agaricus bitorquis</i> pEM, <i>Claviceps purpurea</i> pCIK1, <i>Neurospora crassa</i> Kalilo; <i>Neurospora intermedia</i> Maranh and maize S-2). <p>Two conserved aspartate and one lysine residue have been shown [2,3] to be part of the active site of T7 polymerase. We have used the regions around the first aspartate and around the lysine as signature patterns for this family of polymerases.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-[LIVM]-x(2)-D-[GA]-[ST]-[AC]-[SN]-[GA]-[LIVMFY]-Q [D is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [LIVMF]-x-R-x(3)-K-x(2)-[LIVMF]-M-[PT]-x(2)-Y [K is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1999 / Text revised. References [1] McAllister W.T., Raskin C.A. Mol. Microbiol. 10:1-6(1993).</p> <p>[2] Maksimova T.G., Mustayev A.A., Zaychikov E.F., Lyakhov D.L., Tunitskaya V.L., Akbarov A.K., Luchin S.V., Rechinsky V.O., Chernov B.K., Kochetkov S.N. Eur. J. Biochem. 195:841-847(1991).</p> <p>[3] Sousa R., Chung Y.J., Rose J.P., Wang B.-C. Nature 364:593-599(1993).</p>
RNA_pol_A		RNA polymerase alpha subunit	<p>Accession number: PF00623 Definition: RNA polymerase alpha subunit Author: Bateman A Alignment method of seed: HMM_built_from_alignment Source of seed members: Pfam-B_3 (release 2.1) Gathering cutoffs: 9 0 Trusted cutoffs: 13.50 2.90 Noise cutoffs: 8.50 8.50 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97066998 Reference Title: Structural modules of the large subunits of RNA polymerase. Reference Title: Introducing archaeobacterial and chloroplast split sites in the beta and beta' subunits of Escherichia coli RNA polymerase. Reference Title: polymerase. Reference Author: Severinov K, Mustaev A, Kukarin A, Muzzin O, Bass I, Darst Reference Author: SA, Goldfarb A; Reference Location: J Biol Chem 1996;271:27969-27974. Database Reference INTERPRO; IPR000722; Database reference: PFAMB; PB003218; Comment: -!- RNA polymerases catalyse the DNA dependent polymerisation Comment: of RNA. Prokaryotes contain a single RNA polymerase</p>

974

			<p>Comment: compared to three in eukaryotes (not including mitochondrial.</p> <p>Comment: and chloroplast polymerases).</p> <p>Comment: -!- Members of this family include:</p> <p>Comment: A subunit from eukaryotes</p> <p>Comment: gamma subunit from cyanobacteria</p> <p>Comment: beta' subunit from eubacteria</p> <p>Comment: A' subunit from archaebacteria</p> <p>Comment: B" from chloroplasts</p> <p>Number of members: 202</p>
RNA_pol_A2		RNA polymerase A/beta'/A" subunit	<p>Accession number: PF01854</p> <p>Definition: RNA polymerase A/beta'/A" subunit</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_288 (release 4.2)</p> <p>Gathering cutoffs: -120 -120</p> <p>Trusted cutoffs: -116.50 -116.50</p> <p>Noise cutoffs: -125.00 -125.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmbuild --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 88335550</p> <p>Reference Title: Relatedness of archaebacterial RNA polymerase core subunits</p> <p>Reference Title: to their eubacterial and eukaryotic equivalents.</p> <p>Reference Author: Berghofer B, Krockel L, Kortner C, Truss M, Schallenberg J,</p> <p>Reference Author: Klein A;</p> <p>Reference Location: Nucleic Acids Res 1988;16:8113-8128.</p> <p>Database Reference: INTERPRO; IPR002879;</p> <p>Database reference: PFAMB; PB000546;</p> <p>Database reference: PFAMB; PB000846;</p> <p>Database reference: PFAMB; PB000984;</p> <p>Database reference: PFAMB; PB001168;</p> <p>Comment: RNA polymerases catalyse the DNA dependent polymerisation</p> <p>Comment: of RNA. Prokaryotes contain a single RNA polymerase compared to three in eukaryotes (not including mitochondrial.</p> <p>Comment: and chloroplast polymerases).</p> <p>Comment: This family includes a region of about 400 amino acids.</p> <p>Comment: This family includes the whole archaebacterial A' subunit, but only the C terminal region of the A subunit from eukaryotes</p> <p>Comment: and the beta' subunit from eubacteria.</p> <p>Number of members: 105</p>
RNB	PDOC00904	Ribonuclease II family signature	<p>On the basis of sequence similarities, the following bacterial and eukaryotic proteins seem to form a family:</p> <ul style="list-style-type: none"> - Escherichia coli and related bacteria ribonuclease II (EC 3.1.13.1) (RNase II) (gene rnb) [1]. RNase II is an exonuclease involved in mRNA decay. It degrades mRNA by hydrolyzing single-stranded polyribonucleotides processively in the 3' to 5' direction. - Bacterial ribonuclease R [2], a 3'-5' exoribonuclease that participates in an essential cell function. - Yeast protein SSD1 (or SRK1) which is implicated in the control of the cell cycle G1 phase. - Yeast protein DIS3 [3], which binds to ran (GSP1) and enhances the nucleotide-releasing activity of RCC1 on ran. - Fission yeast protein dis3, which is implicated in mitotic control. - Neurospora crassa cyt-4, a mitochondrial protein required for RNA 5' and 3' end processing and splicing. - Yeast protein MSU1, which is involved in mitochondrial biogenesis. - Synechocystis strain PCC 6803 protein zam [4], which control resistance to the carbonic anhydrase inhibitor acetazolamide. - Caenorhabditis elegans hypothetical protein F48E8.6. <p>The size of these proteins range from 644 residues (rnb) to 1250 (SSD1). While their sequence is highly divergent they share a conserved domain in their C-terminal section [5]. It is possible that this domain plays a role in a putative exonuclease function that would be common to all these proteins. We have developed a signature pattern based on the core of this conserved</p>

			<p>domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [HI]-[FYE]-[GSTAM]-[LIVM]-x(4,5)-Y-[STALV]-x-[FWVAC]-[TV]-[SA]-P-[LIVMA]-[RQ]-[KR]-[FY]-x-D-x(3)-[HQ]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update</p> <p>December 1999 / Pattern and text revised.</p> <p>References</p> <p>[1]</p> <p>Zilhao R., Camelo L., Arraiano C.M. Mol. Microbiol. 8:43-51(1993).</p> <p>[2]</p> <p>Cheng Z.-F., Zuo Y., Li Z., Rudd K.E., Deutscher M.P. J. Biol. Chem. 273:14077-14080(1998).</p> <p>[3]</p> <p>Noguchi E., Hayashi N., Azuma Y., Seki T., Nakamura M., Nakashima N., Yanagida M., He X., Mueller U., Sazer S., Nishimoto T. EMBO J. 15:5595-5605(1996).</p> <p>[4]</p> <p>Beuf L., Bedu S., Cami B., Joset F. Plant Mol. Biol. 27:779-788(1995).</p> <p>[5]</p> <p>Mian I.S. Nucleic Acids Res. 25:3187-3195(1997).</p>
RRF		Ribosome recycling factor	<p>Accession number: PF01765</p> <p>Definition: Ribosome recycling factor</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_949 (release 4.2)</p> <p>Gathering cutoffs: -35 -35</p> <p>Trusted cutoffs: -34.90 -34.90</p> <p>Noise cutoffs: -76.20 -76.20</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 94240115</p> <p>Reference Title: Ribosome recycling factor (ribosome releasing factor) is essential for bacterial growth.</p> <p>Reference Author: Janosi L, Shimizu I, Kaji A;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1994;91:4249-4253.</p> <p>Database Reference INTERPRO; IPR002661;</p> <p>Comment: The ribosome recycling factor (RRF / ribosome release factor) dissociates</p> <p>Comment: the ribosome from the mRNA after termination of translation, and is</p> <p>Comment: essential bacterial growth [1]. Thus ribosomes are "recycled" and ready</p> <p>Comment: for another round of protein synthesis.</p> <p>Number of members: 27</p>
rve		Integrase core domain	<p>Accession number: PF00665</p> <p>Definition: Integrase core domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_10 (release 2.1)</p> <p>Gathering cutoffs: 9.3 9.3</p> <p>Trusted cutoffs: 9.30 9.30</p> <p>Noise cutoffs: 9.20 9.20</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95099322</p> <p>Reference Title: Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases</p>

			Reference Title: [see comments]
			Reference Author: Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R,
			Reference Author: Davies DR;
			Reference Location: Science 1994;266:1981-1986.
			Database Reference: SCOP; 2itg; fa; [SCOP-USA][CATH-PDBSUM]
			Database Reference: INTERPRO; IPR001584;
			Database Reference: PDB; 1cxu A; 56; 198;
			Database Reference: PDB; 1vsh ; 54; 199;
			Database Reference: PDB; 1vsi ; 54; 199;
			Database Reference: PDB; 1vsj ; 54; 199;
			Database Reference: PDB; 1cxq A; 53; 198;
			Database Reference: PDB; 1a5v ; 54; 199;
			Database Reference: PDB; 1a5w ; 54; 199;
			Database Reference: PDB; 1a5x ; 54; 199;
			Database Reference: PDB; 1asv ; 54; 199;
			Database Reference: PDB; 1vsm A; 54; 199;
			Database Reference: PDB; 1czb A; 53; 198;
			Database Reference: PDB; 1asw ; 53; 201;
			Database Reference: PDB; 1cz9 A; 59; 197;
			Database Reference: PDB; 1vsk ; 54; 199;
			Database Reference: PDB; 1vsl A; 54; 199;
			Database Reference: PDB; 1asu ; 53; 207;
			Database Reference: PDB; 1c0m A; 53; 213;
			Database Reference: PDB; 1vsd ; 54; 88;
			Database Reference: PDB; 1vse ; 54; 88;
			Database Reference: PDB; 1c1a B; 55; 213;
			Database Reference: PDB; 1c0m B; 54; 213;
			Database Reference: PDB; 1c0m D; 54; 213;
			Database Reference: PDB; 1c1a A; 53; 213;
			Database Reference: PDB; 1c0m C; 53; 213;
			Database Reference: PDB; 1bhl ; 57; 201;
			Database Reference: PDB; 1bi4 B; 57; 201;
			Database Reference: PDB; 1bl3 B; 57; 201;
			Database Reference: PDB; 1b9f A; 56; 201;
			Database Reference: PDB; 1bis B; 56; 201;
			Database Reference: PDB; 1qs4 B; 56; 201;
			Database Reference: PDB; 1qs4 C; 56; 201;
			Database Reference: PDB; 1biz A; 54; 201;
			Database Reference: PDB; 1itg ; 55; 201;
			Database Reference: PDB; 1bi4 C; 53; 201;
			Database Reference: PDB; 1bl3 C; 53; 201;
			Database Reference: PDB; 2itg ; 53; 201;
			Database Reference: PDB; 1b9d A; 57; 189;
			Database Reference: PDB; 1bi4 A; 57; 201;
			Database Reference: PDB; 1bl3 A; 57; 201;
			Database Reference: PDB; 1bis A; 56; 201;
			Database Reference: PDB; 1biu A; 56; 201;
			Database Reference: PDB; 1biu B; 56; 201;
			Database Reference: PDB; 1biu C; 56; 201;
			Database Reference: PDB; 1qs4 A; 56; 201;
			Database Reference: PDB; 1b92 A; 56; 201;
			Database Reference: PDB; 1biz B; 58; 201;
			Database Reference: PDB; 1b9d A; 382; 390;
			Database Reference: PDB; 1wjb A; 53; 55;
			Database Reference: PDB; 1wjb B; 53; 55;
			Database Reference: PDB; 1wjd A; 53; 55;
			Database Reference: PDB; 1wjd B; 53; 55;
			Database Reference: PDB; 1wjf A; 53; 55;
			Database Reference: PDB; 1wjf B; 53; 55;
			Database reference: PFAMB; PB000048;
			Database reference: PFAMB; PB007709;
			Database reference: PFAMB; PB013923;
			Database reference: PFAMB; PB013938;
			Database reference: PFAMB; PB018509;
			Database reference: PFAMB; PB020302;
			Database reference: PFAMB; PB025327;
			Database reference: PFAMB; PB028352;
			Database reference: PFAMB; PB032740;
			Database reference: PFAMB; PB040612;
			Database reference: PFAMB; PB040636;
			Database reference: PFAMB; PB040684;
			Database reference: PFAMB; PB040695;
			Database reference: PFAMB; PB040730;

			<p>Database reference: PFAMB; PB040824; Database reference: PFAMB; PB041112; Database reference: PFAMB; PB041143; Database reference: PFAMB; PB041275; Database reference: PFAMB; PB041356; Database reference: PFAMB; PB041375; Database reference: PFAMB; PB041456; Database reference: PFAMB; PB041459; Database reference: PFAMB; PB041522; Database reference: PFAMB; PB041665; Database reference: PFAMB; PB041761; Database reference: PFAMB; PB041816; Database reference: PFAMB; PB041885; Comment: Integrase mediates integration of a DNA copy of the viral genome into the host chromosome. Integrase is composed of Comment: three domains. The amino-terminal domain is a zinc binding Comment: domain Integrase_Zn. This domain is the central catalytic Comment: domain. The carboxyl terminal domain that is a non-specific Comment: DNA binding domain integrase. Comment: The catalytic domain acts as an endonuclease when two Comment: nucleotides are removed from the 3' ends of the blunt-ended Comment: viral DNA made by reverse transcription. This domain also Comment: catalyses the DNA strand transfer reaction of the 3' ends of the viral DNA to the 5' ends of the integration site [1]. Number of members: 1147</p>
S4		S4 domain	<p>Accession number: PF01479 Definition: S4 domain Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Medline:99193178 Gathering cutoffs: 17 17 Trusted cutoffs: 17.20 17.20 Noise cutoffs: 16.70 16.70 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 99193178 Reference Title: Novel predicted RNA-binding domains associated with the Reference Title: translation machinery. Reference Author: Aravind L, Koonin EV; Reference Location: J Mol Evol 1999;48:291-302. Reference Number: [2] Reference Medline: 98372721 Reference Title: The crystal structure of ribosomal protein S4 reveals a two-domain molecule with an extensive RNA-binding surface: Reference Title: one domain shows structural homology to the ETS DNA-binding Reference Title: motif. Reference Author: Davies C, Gerstner RB, Draper DE, Ramakrishnan V, White SW; Reference Location: EMBO J 1998;17:4545-4558. Database Reference: SCOP; 1c06; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002942; Database Reference: PDB; 1c05 A; 51; 98; Database Reference: PDB; 1c06 A; 51; 98; Database Reference: PDB; 1dm9 A; 9; 55; Database Reference: PDB; 1dm9 B; 9; 55; Database reference: PFAMB; PB001751; Database reference: PFAMB; PB041147; Database reference: PFAMB; PB041148; Comment: The S4 domain is a small domain consisting of 60-65 amino acid residues Comment: that was detected in the bacterial ribosomal protein S4, eukaryotic Comment: ribosomal S9, two families of pseudouridine synthases, a novel family Comment: of predicted RNA methylases, a yeast protein containing a</p>

978

			<p>pseudouridine</p> <p>Comment: synthetase and a deaminase domain, bacterial tyrosyl-tRNA synthetases,</p> <p>Comment: and a number of uncharacterized, small proteins that may be involved in</p> <p>Comment: translation regulation [1]. The S4 domain probably mediates binding to</p> <p>Comment: RNA.</p> <p>Number of members: 256</p>
SAA_proteins	PDOC00762	Serum amyloid A proteins signature	<p>The serum amyloid A (SAA) proteins comprise a family of vertebrate proteins that associate predominantly with high density lipoproteins (HDL) [1,2]. The synthesis of certain members of the family is greatly increased (as much as a 1000 fold) in inflammation; thus making SAA a major acute phase reactant. While the major physiological function of SAA is unclear, prolonged elevation of plasma SAA levels, as in chronic inflammation, however, results in a pathological condition, called amyloidosis, which affects the liver, kidney and spleen and which is characterized by the highly insoluble accumulation of SAA in these tissues.</p> <p>SAA are proteins of about 110 amino acid residues. As a signature pattern, we selected the most highly conserved region, which is located in the central part of the sequence.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern A-R-G-N-Y-[ED]-A-x-[QKR]-R-G-x-G-G-x-W-A</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update June 1994 / First entry.</p> <p>References [1] Malle E., Steinmetz A., Raynes J.G. Atherosclerosis 102:131-146(1993).</p> <p>[2] Uhlir C.M., Burgess C.J., Sharp P.M., Whitehead A.S. Genomics 19:228-235(1994).</p>
SAM		SAM domain (Sterile alpha motif)	<p>Accession number: PF00536</p> <p>Definition: SAM domain (Sterile alpha motif)</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: [1],[2]</p> <p>Gathering cutoffs: 11 0</p> <p>Trusted cutoffs: 11.00 3.70</p> <p>Noise cutoffs: 10.90 10.90</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96100659</p> <p>Reference Title: SAM: A novel motif in yeast sterile alpha and Drosophila polyhomeotic proteins</p> <p>Reference Author: Ponting CP;</p> <p>Reference Location: Prot Sci 1995;4:1928-1930.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97160498</p> <p>Reference Title: SAM as a protein interaction domain involved in developmental regulation.</p> <p>Reference Author: Shultz J, Ponting CP, Hofmann K, Bork P;</p> <p>Reference Location: Prot Sci 1997;6:249-253.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 99101382</p> <p>Reference Title: The crystal structure of an Eph receptor SAM domain reveals</p> <p>Reference Title: a mechanism for modular dimerization.</p> <p>Reference Author: Stapleton D, Balan I, Pawson T, Sicheri F;</p> <p>Reference Location: Nat Struct Biol 1999;6:44-49.</p> <p>Database reference: SMART; SAM;</p> <p>Database Reference: SCOP; 1b0x; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR001660;</p>

979

			<p>Database Reference PDB; 1b0x A; 910; 973;</p> <p>Database Reference PDB; 1sgg ; 7; 70;</p> <p>Database Reference PDB; 1b4f A; 7; 71;</p> <p>Database Reference PDB; 1b4f C; 7; 71;</p> <p>Database Reference PDB; 1b4f E; 7; 71;</p> <p>Database Reference PDB; 1b4f D; 7; 71;</p> <p>Database Reference PDB; 1b4f H; 7; 71;</p> <p>Database Reference PDB; 1b4f F; 7; 71;</p> <p>Database Reference PDB; 1b4f G; 7; 71;</p> <p>Database Reference PDB; 1b4f B; 7; 71;</p> <p>Database reference: PFAMB; PB008631;</p> <p>Database reference: PFAMB; PB040678;</p> <p>Database reference: PFAMB; PB041111;</p> <p>Database reference: PFAMB; PB041385;</p> <p>Comment: It has been suggested that SAM is an evolutionarily conserved protein</p> <p>Comment: binding domain that is involved in the regulation of numerous</p> <p>Comment: developmental processes in diverse eukaryotes.</p> <p>Comment: The SAM domain can potentially function as a protein interaction</p> <p>Comment: module through its ability to homo- and heterooligomerise with</p> <p>Comment: other SAM domains.</p> <p>Number of members: 110</p>
SAM_decarb ox		Adenosylmet hionine decarboxyl ase	<p>Accession number: PF01536</p> <p>Definition: Adenosylmethionine decarboxylase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_600 (release 4.0)</p> <p>Gathering cutoffs: 11 11</p> <p>Trusted cutoffs: 17.90 17.90</p> <p>Noise cutoffs: 5.70 5.70</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98098079</p> <p>Reference Title: Cloning, mapping and mutational analysis of the</p> <p>Reference Title: S-adenosylmethionine decarboxylase gene in Drosophila</p> <p>Reference Title: melanogaster.</p> <p>Reference Author: Larsson J, Rasmuson-Lestander A;</p> <p>Reference Location: Mol Gen Genet 1997;256:652-660.</p> <p>Database Reference: SCOP; 1jen; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference INTERPRO; IPR001985;</p> <p>Database Reference PDB; 1jen C; 69; 328;</p> <p>Database Reference PDB; 1jen A; 69; 329;</p> <p>Database Reference PDB; 1jen B; 4; 67;</p> <p>Database Reference PDB; 1jen D; 5; 66;</p> <p>Comment: This is a family of S-adenosylmethionine decarboxylase (SAMDC) proenzymes.</p> <p>Comment: In the biosynthesis of polyamines SAMDC produces decarboxylated</p> <p>Comment: S-adenosylmethionine, which serves as the aminopropyl moiety necessary</p> <p>Comment: for spermidine and spermine biosynthesis from putrescine [1]. The Pfam</p> <p>Comment: alignment contains both the alpha and beta chains that are cleaved to</p> <p>Comment: form the active enzyme.</p> <p>Number of members: 34</p>
SBF		Sodium Bile acid symporter family	<p>Accession number: PF01758</p> <p>Definition: Sodium Bile acid symporter family</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_697 (release 4.2)</p> <p>Gathering cutoffs: -19 -19</p> <p>Trusted cutoffs: -12.50 -12.50</p> <p>Noise cutoffs: -26.40 -26.40</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97377989</p>

980

			<p>Reference Title: Isolation of three contiguous genes, ACR1, ACR2 and ACR3, Reference Title: involved in resistance to arsenic compounds in the yeast Reference Title: <i>Saccharomyces cerevisiae</i>. Reference Author: Bobrowicz P, Wysocki R, Owsianik G, Goffeau A, Ulaszewski Reference Author: S; Reference Location: Yeast 1997;13:819-828. Reference Number: [2] Reference Medline: 92073340 Reference Title: Functional expression cloning and characterization of the Reference Title: hepatocyte Na⁺/bile acid cotransport system. Reference Author: Hagenbuch B, Stieger B, Foguet M, Lubbert H, Meier PJ; Reference Location: Proc Natl Acad Sci U S A 1991;88:10629-10633. Database Reference: INTERPRO; IPR002657; Database reference: PFAMB; PB041594; Comment: This family consists of Na⁺/bile acid co-transporters. Comment: These transmembrane proteins function in the liver Comment: in the uptake of bile acids from portal blood plasma Comment: a process mediated by the co-transport of Na⁺ [2]. Comment: Also in the family is ARC3 from <i>S. cerevisiae</i> Swiss:Q06598 Comment: this is a putative transmembrane protein involved in Comment: resistance to arsenic compounds [1]. Number of members: 29</p>
Sec7		Sec7 domain	<p>Accession number: PF01369 Definition: Sec7 domain Author: Bateman A Alignment method of seed: Clustalw_manual Source of seed members: Pfam-B_1629 (release 3.0) Gathering cutoffs: 25 25 Trusted cutoffs: 101.50 101.50 Noise cutoffs: 13.20 13.20 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 98169075 Reference Title: Structure of the Sec7 domain of the Arf exchange factor Reference Title: ARNO. Reference Author: Cherfils J, Menetrey J, Mathieu M, Le Bras G, Robineau S, Reference Author: Beraud-Dufour S, Antonny B, Chardin P; Reference Location: Nature 1998;392:101-105. Reference Number: [2] Reference Medline: 97100951 Reference Title: A human exchange factor for ARF contains Sec7- and Reference Title: pleckstrin- homology domains. Reference Author: Chardin P, Paris S, Antonny B, Robineau S, Beraud-Dufour S, Reference Author: Jackson CL, Chabre M Reference Location: Nature 1996;384:481-484. Database Reference: SCOP; 1pbv; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR000904; Database Reference: PDB; 1pbv ; 58; 243; Database Reference: PDB; 1bc9 ; 59; 244; Comment: The Sec7 domain is a guanine-nucleotide-exchange-factor Comment: (GEF) Comment: for the arf family [2]. Number of members: 32</p>
Seedstore_2 S		2S seed storage family	<p>Accession number: PF01631 Definition: 2S seed storage family Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1154 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 95.10 95.10 Noise cutoffs: -0.20 10.10 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 97121264</p>

981

			<p>Reference Title: 1H NMR assignment and global fold of napin Bnlb, a representative 2S albumin seed protein.</p> <p>Reference Author: Rico M, Bruix M, Gonzalez C, Monsalve RI, Rodriguez R;</p> <p>Reference Location: Biochemistry 1996;35:15672-15682.</p> <p>Database Reference: SCOP; 1pnb; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR000617;</p> <p>Database reference: PFAMB; PB029622;</p> <p>Comment: Members of this family are composed of two chains (both included in</p> <p>Comment: the alignment), these are co-translated and later cleaved.</p> <p>The two</p> <p>Comment: chains are disulphide linked together.</p> <p>Number of members: 27</p>
SH2	PDOC50001	Src homology 2 (SH2) domain profile	<p>The Src homology 2 (SH2) domain is a protein domain of about 100 amino-acid residues first identified as a conserved sequence region between the oncoproteins Src and Fps [1]. Similar sequences were later found in many other intracellular signal-transducing proteins [2]. SH2 domains function as regulatory modules of intracellular signalling cascades by interacting with high affinity to phosphotyrosine-containing target peptides in a sequence-specific and strictly phosphorylation-dependent manner [3,4,5,6].</p> <p>The SH2 domain has a conserved 3D structure consisting of two alpha helices and six to seven beta-strands. The core of the domain is formed by a continuous beta-meander composed of two connected beta-sheets [7].</p> <p>So far, SH2 domains have been identified in the following proteins:</p> <ul style="list-style-type: none"> - Many vertebrate, invertebrate and retroviral cytoplasmic (non-receptor) protein tyrosine kinases. In particular in the Src, Abl, Btk, Csk and ZAP70 families of kinases. - Mammalian phosphatidylinositol-specific phospholipase C gamma-1 and -2. <p>Two</p> <p>copies of the SH2 domain are found in those proteins in between the catalytic 'X-' and 'Y-boxes' (see <PDOC50007>).</p> <ul style="list-style-type: none"> - Mammalian phosphatidylinositol 3-kinase regulatory p85 subunit. - Some vertebrate and invertebrate protein-tyrosine phosphatases. - Mammalian Ras GTPase-activating protein (GAP). - Adaptor proteins mediating binding of guanine nucleotide exchange factors to growth factor receptors: vertebrate GRB2, Caenorhabditis elegans sem-5 and Drosophila DRK. - Mammalian Vav oncoprotein, a guanine-nucleotide exchange factor of the CDC24 family. - Miscellaneous proteins interacting with vertebrate receptor protein tyrosine kinases: oncoprotein Crk, mammalian cytoplasmic proteins Nck, Shc. - STAT proteins (signal transducers and activators of transcription). - Chicken tensin. - Yeast transcriptional control protein SPT6. <p>The profile developed to detect SH2 domains is based on a structural alignment consisting of 8 gap-free blocks and 7 linker regions totaling 92 match positions.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Sequences known to belong to this class detected by the profile ALL.</p> <p>Other sequence(s) detected in SWISS-PROT protein tyrosine kinases JAK1 and JAK2.</p> <p>Expert(s) to contact by email Zvelebil M. marketa@ludwig.ucl.ac.uk</p> <p>Last update November 1995 / First entry.</p> <p>References</p> <p>[1] Sadowski J., Stone J.C., Pawson T. Mol. Cell. Biol. 6:4396-4408(1986).</p> <p>[2] Russel R.B., Breed J., Barton G.J.</p>

			<p>FEBS Lett. 304:15-20(1992).</p> <p>[3] Marangere L.E.M., Pawson T. J. Cell Sci. Suppl. 18:97-104(1994).</p> <p>[4] Pawson T., Schlessinger J. Curr. Biol. 3:434-442(1993).</p> <p>[5] Mayer B.J., Baltimore D. Trends Cell. Biol. 3:8-13(1993).</p> <p>[6] Pawson T. Nature 373:573-580(1995).</p> <p>[7] Kuriyan J., Cowburn D. Curr. Opin. Struct. Biol. 3:828-837(1993).</p>
Shikimate_D H		Shikimate / quininate 5- dehydrogena se	<p>Accession number: PF01488</p> <p>Definition: Shikimate / quinate 5-dehydrogenase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_336 (release 4.0)</p> <p>Gathering cutoffs: -50 -50</p> <p>Trusted cutoffs: -48.00 -48.00</p> <p>Noise cutoffs: -82.00 -82.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96048023</p> <p>Reference Title: The molecular biology of multidomain proteins. Selected examples.</p> <p>Reference Author: Hawkins AR, Lamb HK;</p> <p>Reference Location: Eur J Biochem 1995;232:7-18.</p> <p>Database Reference: INTERPRO; IPR002907;</p> <p>Comment: This family contains both shikimate and quinate dehydrogenases.</p> <p>Comment: Shikimate 5-dehydrogenase catalyses the conversion of shikimate to 5-dehydroshikimate. This reaction is part of the shikimate pathway which is involved in the biosynthesis of aromatic amino acids.</p> <p>Comment: Quinate 5-dehydrogenase catalyses the conversion of quinate to 5-dehydroquininate. This reaction is part of the quinate pathway where quinic acid is exploited as a source of carbon in prokaryotes and microbial eukaryotes.</p> <p>Comment: Both the shikimate and quinate pathways share two common pathway metabolites 3-dehydroquininate and dehydroshikimate.</p> <p>Number of members: 58</p>
Sigma54_fact ors	PDOC00593	Sigma-54 factors family signatures and profile	<p>Sigma factors [1] are bacterial transcription initiation factors that promote the attachment of the core RNA polymerase to specific initiation sites and are then released. They alter the specificity of promoter recognition. Most bacteria express a multiplicity of sigma factors. Two of these factors, sigma-70 (gene rpoD), generally known as the major or primary sigma factor, and sigma-54 (gene rpoN or ntrA) direct the transcription of a wide variety of genes. The other sigma factors, known as alternative sigma factors, are required for the transcription of specific subsets of genes.</p> <p>With regard to sequence similarity, sigma factors can be grouped into two classes: the sigma-54 and sigma-70 families. The sigma-70 family has many different sigma factors (see the relevant entry <PDOC00592>). The sigma-54 family consists exclusively of sigma-54 factor [2,3] required for the transcription of promoters that have a characteristic -24 and -12 consensus recognition element but which are devoid of the typical -10,-35 sequences recognized by the major sigma factors. The sigma-54 factor is also characterized by its interaction with ATP-dependent positive regulatory</p>

			<p>proteins that bind to upstream activating sequences.</p> <p>Structurally sigma-54 factors consist of three distinct regions:</p> <ul style="list-style-type: none"> - A relatively well conserved N-terminal glutamine-rich region of about 50 residues that contains a potential leucine zipper motif. - A region of variable length which is not well conserved. - A well conserved C-terminal region of about 350 residues that contains a second potential leucine zipper, a potential DNA-binding 'helix-turn-helix' motif and a perfectly conserved octapeptide whose function is not known. <p>We developed two signature patterns for this family of sigma factors. The first starts two residues before the N-terminal extremity of the helix-turn-helix region and ends two residues before its C-terminal extremity. The second is the conserved octapeptide. A profile has also been designed that covers the whole C-terminal region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-[LIVM]-x-[LIVM]-x(2)-[LIVM]-A-x(2)-[LIVMFT]-x(2)-[HS]-x-S-T-[LIVM]-S-R Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern R-R-T-[IV]-[ATN]-K-Y-R Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.</p> <p>Last update July 1999 / Patterns and text revised.</p> <p>References [1] Hermann J.D., Chamberlin M.J. Annu. Rev. Biochem. 57:839-872(1988).</p> <p>[2] Thoeny B., Hennecke H. FEMS Microbiol. Rev. 5:341-358(1989).</p> <p>[3] Merrick M.J. Mol. Microbiol. 10:903-909(1993).</p>
SLH	PDOC00823	S-layer homology domain signature	<p>S-layers are paracrystalline mono-layered assemblies of (glyco)proteins which coat the surface of bacteria [1]. Several S-layer proteins and some other cell wall proteins contain one or more copies of a domain of about 50-60 residues, which has been called SLH (for S-layer homology) [2]. There is strong evidence that this domain serves as an anchor to the peptidoglycan [3]. The SLH domain has been found in:</p> <ul style="list-style-type: none"> - S-layer glycoprotein of <i>Acetogenium kivui</i> (3 copies). - S-layer 125 Kd protein of <i>Bacillus sphaericus</i> (3 copies). - S-layer protein of <i>Bacillus anthracis</i> (3 copies). - S-layer protein of <i>Bacillus licheniformis</i> (3 copies). - S-layer protein (HWP) from <i>Bacillus brevis</i> strain HPD31 (3 copies). - Middle cell wall protein (MWP) from <i>Bacillus brevis</i> strain 47 (3 copies). - S-layer protein (p100) of <i>Thermus thermophilus</i> (1 copy). - Outer membrane protein Omp-alpha from <i>Thermotoga maritima</i> (1 copy). - Cellulosome anchoring protein (gene ancA), outer layer protein B (OlpB) and a further potential cell surface glycoprotein from <i>Clostridium thermocellum</i> (3 copies; the first copy is missing its N-terminal third which is appended to the end of the third copy; may have arisen by circular permutation). - Amylopullulanase (gene amyB) from <i>Thermoanaerobacter thermosulfurogenes</i> (3 copies) - Amylopullulanase (gene aapT) from <i>Bacillus</i> strain XAL-601 (3 copies).

			<ul style="list-style-type: none"> - Endoglucanase from Bacillus strain KSM-635 (3 copies). - Exoglucanase (gene xynX) from Clostridium thermocellum (3 copies). - Xylanase A (gene xynA) from Thermoanaerobacter saccharolyticum (2 copies; 3 copies if a frameshift is taken into account). - Protein involved in butirosin production (ButB) from Bacillus circulans (2 incomplete copies; 3 copies if three frameshifts are taken into account). - Two hypothetical proteins from Synechocystis strain PCC 6803 (1 copy each). - A hypothetical protein with sequence similarity to amylopullulanases found 3' of amylase gene from Bacillus circulans (fragment of 1 copy; 3 copies if two frameshifts are taken into account). <p>SLH domains are found at the N- or C-termini of mature proteins. They occur in single copy followed by a predicted coiled coil domain, or in three contiguous copies. Structurally, the SLH domain is predicted to contain two alpha-helices flanking a beta strand. The SLH sequences are fairly divergent with an average identity of about 25%. It is however possible to build a sequence pattern that starts at the second position of the domain and that spans 3/4 of its length.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LVFYT]-x-[DA]-x(2,5)-[DNQSATPHY]-[FYWPDA]-x(4)-[LIV]-x(2)-[GTALV]-x(4,6)-[LIVFYC]-x(2)-G-x-[PGSTA]-x(2,3)-[MFYA]-x-[PGAV]-x(3,10)-[LIVMA]-[STKR]-[RY]-x-[EQ]-x-[STALIVM]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Expert(s) to contact by email Lupas A.N. lupas@vms.biochem.mpg.de</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Beveridge T.J. Curr. Opin. Struct. Biol. 4:204-212(1994).</p> <p>[2] Lupas A., Engelhardt H., Peters J., Santarius U., Volker S., Baumeister W. J. Bacteriol. 176:1224-1233(1994).</p> <p>[3] Lemaire M., Ohayon H., Gounon P., Fujino T., Beguin P. J. Bacteriol. 177:2451-2459(1995).</p>
Smr		Smr domain	<p>Accession number: PF01713</p> <p>Definition: Smr domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 0 0</p> <p>Trusted cutoffs: 1.40 1.40</p> <p>Noise cutoffs: -7.90 -7.90</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 10431172</p> <p>Reference Title: Smr: a bacterial and eukaryotic homologue of the C-terminal</p> <p>Reference Title: region of the MutS2 family.</p> <p>Reference Author: Moreira D, Philippe H;</p> <p>Reference Location: Trends Biochem Sci 1999;24:298-300.</p> <p>Database Reference: INTERPRO; IPR002625;</p> <p>Comment: This family includes the Smr (Small MutS Related) proteins,</p> <p>Comment: and the C-terminal region of the MutS2 protein. It has been</p> <p>Comment: suggested that this domain interacts with the MutS1</p> <p>Comment: Swiss:P23909 protein in the case of Smr proteins and with</p> <p>Comment: the N-terminal MutS related region of MutS2</p> <p>Swiss:P94545 [1].</p> <p>Number of members: 14</p>

985

SRF-TF	PDOC00302	MADS-box domain signature and profile	<p>A number of transcription factors contain a conserved domain of 56 amino-acid residues, sometimes known as the MADS-box domain [E1]. They are listed below:</p> <ul style="list-style-type: none"> - Serum response factor (SRF) [1], a mammalian transcription factor that binds to the Serum Response Element (SRE). This is a short sequence of dyad symmetry located 300 bp to the 5' end of the transcription initiation site of genes such as c-fos. - Mammalian myocyte-specific enhancer factors 2A to 2D (MEF2A to MEF2D). These proteins are transcription factor which binds specifically to the MEF2 element present in the regulatory regions of many muscle-specific genes. - Drosophila myocyte-specific enhancer factor 2 (MEF2). - Yeast GRM/PRTF protein (gene MCM1) [2], a transcriptional regulator of mating-type-specific genes. - Yeast arginine metabolism regulation protein I (gene ARGR1 or ARG80). - Yeast transcription factor RLM1. - Yeast transcription factor SMP1. - Arabidopsis thaliana agamous protein (AG) [3], a probable transcription factor involved in regulating genes that determines stamen and carpel development in wild-type flowers. Mutations in the AG gene result in the replacement of the stamens by petals and the carpels by a new flower. - Arabidopsis thaliana homeotic proteins Apetala1 (AP1), Apetala3 (AP3) and Pistillata (PI) which act locally to specify the identity of the floral meristem and to determine sepal and petal development [4]. - Antirrhinum majus and tobacco homeotic protein deficiens (DEFA) and globosa (GLO) [5]. Both proteins are transcription factors involved in the genetic control of flower development. Mutations in DEFA or GLO cause the transformation of petals into sepals and of stamens into carpels. - Arabidopsis thaliana putative transcription factors AGL1 to AGL6 [6]. - Antirrhinum majus morphogenetic protein DEF H33 (squamosa). <p>In SRF, the conserved domain has been shown [1] to be involved in DNA-binding and dimerization. We have derived a pattern that spans the complete length of the domain. The profile also spans the length of the MADS-box.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern R-x-[RK]-x(5)-I-x-[DNGSK]-x(3)-[KR]-x(2)-T-[FY]-x-[RK](3)-x(2)-[LIVM]-x-K(2)-A-x-E-[LIVM]-[STA]-x-L-x(4)-[LIVM]-x-[LIVM](3)-x(6)-[LIVMF]-x(2)-[FY]</p> <p>Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this documentation entry is linked to both signature patterns and a profile. As the profile is much more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so.</p> <p>Last update July 1999 / Pattern and text revised.</p> <p>References [1] Norman C., Runswick M., Pollock R., Treisman R. Cell 55:989-1003(1988).</p> <p>[2] Passmore S., Maine G.T., Elble R., Christ C., Tye B.-K. J. Mol. Biol. 204:593-606(1988).</p> <p>[3] Yanofsky M., Ma H., Bowman J., Drews G., Feldmann K.A., Meyerowitz E.M. Nature 346:35-39(1990).</p> <p>[4] Goto K., Meyerowitz E.M. Genes Dev. 8:1548-1560(1994).</p>
--------	-----------	---------------------------------------	--

			<p>[5] Troebner W., Ramirez L., Motte P., Hue I., Huijser P., Loennig W.-E., Saedler H., Sommer H., Schwartz-Sommer Z. EMBO J. 11:4693-4704(1992).</p> <p>[6] Ma H., Yanofsky M.F., Meyerowitz E.M. Genes Dev. 5:484-495(1991).</p> <p>[E1] http://transfac.gbf-braunschweig.de/cgi-bin/qt/getEntry.pl?C0014</p>
SRP19		SRP19 protein	<p>Accession number: PF01922 Definition: SRP19 protein Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 31.20 31.20 Noise cutoffs: -28.50 -28.50 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 89041541 Reference Title: Isolation and characterization of a cDNA clone encoding the Reference Title: 19 kDa protein of signal recognition particle (SRP): Reference Title: expression and binding to 7SL RNA. Reference Author: Lingelbach K, Zwieb C, Webb JR, Marshallsay C, Hoben PJ, Reference Author: Walter P, Dobberstein B; Reference Location: Nucleic Acids Res 1988;16:9431-9442. Reference Number: [2] Reference Medline: 92220168 Reference Title: SEC65 gene product is a subunit of the yeast signal Reference Title: recognition particle required for its integrity. Reference Author: Hann BC, Stirling CJ, Walter P; Reference Location: Nature 1992;356:532-533. Reference Number: [3] Reference Medline: 92220169 Reference Title: The S. cerevisiae SEC65 gene encodes a component of yeast Reference Title: signal recognition particle with homology to human SRP19. Reference Author: Stirling CJ, Hewitt EW; Reference Location: Nature 1992;356:534-537. Database Reference: INTERPRO; IPR002778; Comment: The signal recognition particle (SRP) binds to the signal peptide of Comment: proteins as they are being translated. The binding of the SRP halts Comment: translation and the complex is then transported to the endoplasmic Comment: reticulum's cytoplasmic surface. The SRP then aids translocation of Comment: the protein through the ER membrane. The SRP is a ribonucleoprotein Comment: that is composed of a small RNA and several proteins. One of these Comment: proteins is the SRP19 protein [1] (Sec65 in yeast [2,3]). Number of members: 13</p>
SSB	PDOC00602	Single-strand binding protein family signatures	<p>The Escherichia coli single-strand binding protein [1] (gene ssb), also known as the helix-destabilizing protein, is a protein of 177 amino acids. It binds tightly, as a homotetramer, to single-stranded DNA (ss-DNA) and plays an important role in DNA replication, recombination and repair.</p> <p>Closely related variants of SSB are encoded in the genome of a variety of large self-transmissible plasmids. SSB has also been characterized in bacteria such as Proteus mirabilis or Serratia marcescens.</p> <p>Eukaryotic mitochondrial proteins that bind ss-DNA and are probably involved in mitochondrial DNA replication are structurally and evolutionary related to</p>

			<p>prokaryotic SSB. Proteins currently known to belong to this subfamily are listed below [2].</p> <ul style="list-style-type: none"> - Mammalian protein Mt-SSB (P16). - Xenopus Mt-SSBs and Mt-SSBr. - Drosophila MtSSB. - Yeast protein RIM1. <p>We have developed two signature patterns for these proteins. The first is a conserved region in the N-terminal section of the SSB's. The second is a centrally located region which, in <i>Escherichia coli</i> SSB, is known to be involved in the binding of DNA.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMF]-[NST]-[KRHST]-[LIVM]-x-[LIVMF](2)-G-[NHRK]-[LIVMA]-[GST]-x-[DENT] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern T-x-W-[HY]-[RNS]-[LIVM]-x-[LIVMF]-[FY]-[NGKR] Sequences known to belong to this class detected by the pattern A majority. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update December 1999 / Patterns and text revised.</p> <p>References [1] Meyer R.R., Laine P.S. Microbiol. Rev. 54:342-380(1990).</p> <p>[2] Stroumbakis N.D., Li Z., Tolias P.P. Gene 143:171-177(1994).</p>
START		START domain	<p>Accession number: PF01852 Definition: START domain Author: SMART Alignment method of seed: Manual Source of seed members: Alignment kindly provided by SMART Gathering cutoffs: 25 25 Trusted cutoffs: 106.20 106.20 Noise cutoffs: -20.90 -20.90 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 99257451 Reference Title: START: a lipid-binding domain in Star, HD-ZIP and signalling proteins. Reference Author: Ponting CP, Aravind L; Reference Location: Trends Biochem Sci 1999;24:130-132. Database reference: SMART; START; Database Reference: INTERPRO; IPR002913; Number of members: 41</p>
Sterol_desat		Sterol desaturase	<p>Accession number: PF01598 Definition: Sterol desaturase Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_905 (release 4.1) Gathering cutoffs: -13 -13 Trusted cutoffs: 12.90 12.90 Noise cutoffs: -44.50 -44.50 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 91323727 Reference Title: Cloning, disruption and sequence of the gene encoding yeast Reference Title: C-5 sterol desaturase. Reference Author: Arthington BA, Bennett LG, Skatrud PL, Guynn CJ, Barbuch Reference Author: RJ, Ulbright CE, Bard M;</p>

			<p>Reference Location: Gene 1991;102:39-44. Reference Number: [2] Reference Medline: 96133902 Reference Title: Cloning and characterization of ERG25, the <i>Saccharomyces cerevisiae</i> gene encoding C-4 sterol methyl oxidase. Reference Author: Bard M, Bruner DA, Pierson CA, Lees ND, Biermann B, Frye L, Reference Author: Koegel C, Barbuch R; Reference Location: Proc Natl Acad Sci U S A 1996;93:186-190. Reference Number: [3] Reference Medline: 96351930 Reference Title: Molecular characterization of the CER1 gene of <i>arabidopsis</i> Reference Title: involved in epicuticular wax biosynthesis and pollen fertility. Reference Author: Aarts MG, Keijzer CJ, Stiekema WJ, Pereira A; Reference Location: Plant Cell 1995;7:2115-2127. Database Reference: INTERPRO; IPR001541; Database reference: PFAMB; PB041851; Comment: This family includes C-5 sterol desaturase and C-4 sterol methyl Comment: oxidase. Members of this family are involved in Comment: cholesterol biosynthesis Comment: and biosynthesis a plant cuticular wax. These enzymes Comment: contain many Comment: conserved histidine residues. Members of this family are Comment: integral Comment: membrane proteins. Number of members: 34</p>
Sulfate_trans p	PDOC00870	Sulfate transporters signature	<p>A number of proteins involved in the transport of sulfate across a membrane as well as some yet uncharacterized proteins have been shown [1,2] to be evolutionary related. These proteins are:</p> <ul style="list-style-type: none"> - <i>Neurospora crassa</i> sulfate permease II (gene <i>cys-14</i>). - Yeast sulfate permeases (genes <i>SUL1</i> and <i>SUL2</i>). - Rat sulfate anion transporter 1 (<i>SAT-1</i>). - Mammalian DTDST, a probable sulfate transporter which, in Human, is involved in the genetic disease, diastrophic dysplasia (DTD). - Sulfate transporters 1, 2 and 3 from the legume <i>Stylosanthes hamata</i>. - Human pendrin (gene <i>PDS</i>), which is involved in a number of hearing loss genetic diseases. - Human protein DRA (Down-Regulated in Adenoma). - Soybean early nodulin 70. - <i>Escherichia coli</i> hypothetical protein <i>ychM</i>. - <i>Caenorhabditis elegans</i> hypothetical protein F41D9.5. <p>As expected by their transport function, these proteins are highly hydrophobic and seem to contain about 12 transmembrane domains. The best conserved region seems to be located in the second transmembrane region and is used as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [PAV]-x-Y-[GS]-L-Y-[STAG]{2}-x(4)-[LIVFYA]-[LIVST]-[YI]-x(3)-[GA]-[GST]-S-[KR] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update July 1999 / Pattern and text revised. References [1] Sandal N.N., Marcker K.A. Trends Biochem. Sci. 19:19-19(1994). [2] Smith F.W., Hawkesford M.J., Prosser I.M., Clarkson D.T. Mol. Gen. Genet. 247:709-715(1995).</p>

989

Synuclein		Synuclein	<p>Accession number: PF01387</p> <p>Definition: Synuclein</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 197.80 197.80</p> <p>Noise cutoffs: -33.80 -33.80</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98424410</p> <p>Reference Title: The synuclein family.</p> <p>Reference Author: Lavedan C;</p> <p>Reference Location: Genome Res 1998;8:871-880.</p> <p>Database Reference INTERPRO; IPR001058;</p> <p>Comment: There are three types of synucleins in humans, these</p> <p>Comment: are called alpha, beta and gamma. Alpha synuclein has</p> <p>Comment: been found mutated in families with autosomal dominant</p> <p>Comment: Parkinson's disease. A peptide of alpha synuclein has</p> <p>Comment: also been found in amyloid plaques in Alzheimer's</p> <p>Comment: patients.</p> <p>Number of members: 12</p>
TEA	PDOC00479	TEA domain signature	<p>The TEA domain [1,E1] is a DNA-binding region of about 66 to 68 amino acids which has been found in the N-terminal section of the following nuclear regulatory proteins:</p> <ul style="list-style-type: none"> - Mammalian enhancer factor TEF-1. TEF-1 can bind to two distinct sequences in the SV40 enhancer and is a transcriptional activator. - Mammalian TEF-3, TEF-4 and TEF-5 [2], putative transcriptional activators highly similar to TEF-1. - Drosophila scalloped protein (gene sd), a probable transcription factor that functions in the regulation of cell-specific gene expression during Drosophila development, particularly in the differentiation of the nervous system [3]. - Emericella nidulans regulatory protein abaA. AbaA is involved in the regulation of conidiation (asexual spore); its expression leads to the cessation of vegetative growth. - Yeast trans-acting factor TEC1. TEC1 is involved in the activation of the Ty1 retrotransposon. - Caenorhabditis elegans hypothetical protein F28B12.2. <p>As a signature pattern, we have used positions 39 to 67 of the TEA domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-R-N-E-L-I-x(2)-Y-I-x(3)-[TC]-x(3)-R-T-[RK](2)-Q-[LIVM]-S-S-H-[LIVM]-Q-V</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References</p> <p>[1]</p> <p>Buerglin T.R.</p> <p>Cell 66:11-12(1991).</p> <p>[2]</p> <p>Jacquemin P., Hwang J.-J., Martial J.A., Dolle P., Davidson I.</p> <p>J. Biol. Chem. 271:21775-21785(1996).</p> <p>[3]</p> <p>Campbell S.D., Inamdar M., Rodrigues V., Raghavan V., Palazzolo M., Chovnick A.</p> <p>Genes Dev. 6:367-379(1992).</p> <p>[E1]</p> <p>http://transfac.gbf-braunschweig.de/cgi-bin/qt/getEntry.pl?C0024</p>
TGT		Queuine	<p>Accession number: PF01702</p>

		tRNA-ribosyltransferase	<p>Definition: Queuine tRNA-ribosyltransferase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1643 (release 4.1)</p> <p>Gathering cutoffs: -132 -132</p> <p>Trusted cutoffs: -110.00 -110.00</p> <p>Noise cutoffs: -155.40 -155.40</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96256303</p> <p>Reference Title: Crystal structure of tRNA-guanine transglycosylase: RNA modification by base exchange.</p> <p>Reference Author: Romier C, Reuter K, Suck D, Ficner R;</p> <p>Reference Location: EMBO J 1996;15:2850-2857.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 93287116</p> <p>Reference Title: tRNA-guanine transglycosylase from Escherichia coli.</p> <p>Reference Title: Overexpression, purification and quaternary structure.</p> <p>Reference Author: Garcia GA, Koch KA, Chong S;</p> <p>Reference Location: J Mol Biol 1993;231:489-497.</p> <p>Database Reference: SCOP; 1pud; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002616;</p> <p>Database Reference: PDB; 1efz A; 138; 379;</p> <p>Database Reference: PDB; 1enu A; 138; 379;</p> <p>Database Reference: PDB; 1pud ; 138; 379;</p> <p>Database Reference: PDB; 1wkd ; 138; 379;</p> <p>Database Reference: PDB; 1wke ; 138; 379;</p> <p>Database Reference: PDB; 1wkf ; 138; 379;</p> <p>Database reference: PFAMB; PB037884;</p> <p>Comment: This is a family of queuine tRNA-ribosyltransferases</p> <p>Comment: EC:2.4.2.29, also known as tRNA-guanine transglycosylase</p> <p>Comment: and guanine insertion enzyme.</p> <p>Comment: Queuine tRNA-ribosyltransferase modifies tRNAs for asparagine,</p> <p>Comment: aspartic acid, histidine and tyrosine with queuine.</p> <p>Comment: It catalyses the exchange of guanine-34 at the wobble position with</p> <p>Comment: 7-aminomethyl-7-deazaguanine, and the addition of a cyclopentenediol</p> <p>Comment: moiety to 7-aminomethyl-7-deazaguanine-34 tRNA; giving a hypermodified</p> <p>Comment: base queuine in the wobble position [1,2].</p> <p>Comment: The aligned region contains a zinc binding motif C-x-C-x2-C-x29-H,</p> <p>Comment: and important tRNA and 7-aminomethyl-7-deazaguanine binding residues [1].</p> <p>Number of members: 24</p>
Thi4		Thi4 family	<p>Accession number: PF01946</p> <p>Definition: Thi4 family</p> <p>Author: Enright A, Ouzounis C, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Enright A</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 526.80 526.80</p> <p>Noise cutoffs: -105.00 -105.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95050223</p> <p>Reference Title: Cloning, nucleotide sequence, and regulation of</p> <p>Reference Title: Schizosaccharomyces pombe thi4, a thiamine biosynthetic</p> <p>Reference Title: gene.</p> <p>Reference Author: Zurlinden A, Schweingruber ME;</p> <p>Reference Location: J Bacteriol 1994;176:6631-6635.</p> <p>Database Reference: INTERPRO; IPR002922;</p> <p>Comment: This family includes Swiss:P32318 a putative thiamine biosynthetic</p> <p>Comment: enzyme.</p> <p>Number of members: 14</p>

991

ThiC		ThiC family	<p>Accession number: PF01964 Definition: ThiC family Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 1047.20 1047.20 Noise cutoffs: -338.20 -338.20 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 93163063 Reference Title: Structural genes for thiamine biosynthetic enzymes (thiCEFGH) in Escherichia coli K-12. Reference Author: Vander Horn PB, Backstrom AD, Stewart V, Begley TP; Reference Location: J Bacteriol 1993;175:982-992. Reference Number: [2] Reference Medline: 99311269 Reference Title: Thiamin biosynthesis in prokaryotes. Reference Author: Begley TP, Downs DM, Ealick SE, McLafferty FW, Van Loon AP, Reference Author: Taylor S, Campobasso N, Chiu HJ, Kinsland C, Reddick JJ, Xi Reference Author: J; Reference Location: Arch Microbiol 1999;171:293-300. Reference Number: [3] Reference Medline: 97284509 Reference Title: Characterization of the Bacillus subtilis thiC operon involved in thiamine biosynthesis. Reference Author: Zhang Y, Taylor SV, Chiu HJ, Begley TP; Reference Location: J Bacteriol 1997;179:3030-3035. Database Reference: INTERPRO; IPR002817; Comment: ThiC is found within the thiamine biosynthesis operon. ThiC is Comment: involved in pyrimidine biosynthesis [2]. Comment: ThiC catalyzes the substitution of the pyrophosphate of 2-methyl-4-amino-5-hydroxymethylpyrimidine pyrophosphate by Comment: 4-methyl-5-(beta-hydroxyethyl)thiazole phosphate to yield thiamine Comment: phosphate [3]. Number of members: 12</p>
ThiJ		ThiJ/Pfpl family	<p>Accession number: PF01965 Definition: ThiJ/Pfpl family Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: -40.2 -40.2 Trusted cutoffs: -40.20 -40.20 Noise cutoffs: -47.00 -47.00 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97039868 Reference Title: The thiJ locus and its relation to phosphorylation of hydroxymethylpyrimidine in Escherichia coli. Reference Author: Mizote T, Tsuda M, Nakazawa T, Nakayama H; Reference Location: Microbiology 1996;142:2969-2974. Reference Number: [2] Reference Medline: 96196168 Reference Title: Sequence, expression in Escherichia coli, and analysis of the gene encoding a novel intracellular protease (Pfpl) from the hyperthermophilic archaeon Pyrococcus furiosus. Reference Author: Halio SB, Blumentals II, Short SA, Merrill BM, Kelly RM; Reference Location: J Bacteriol 1996;178:2605-2612. Database Reference: INTERPRO; IPR002818; Database reference: PFAMB; PB002774; Database reference: PFAMB; PB007213; Database reference: PFAMB; PB041784; Comment: This family includes ThiJ a thiamine biosynthesis enzyme [1] that catalyses the phosphorylation of hydroxymethylpyrimidine (HMP) to HMP monophosphate</p>

			<p>EC:2.7.1.49.</p> <p>Comment: The family also includes a the protease Pfpl</p> <p>Swiss:Q51732 [2].</p> <p>Number of members: 34</p>
Thr_dehydrat_C		C-terminal domain of Threonine dehydratase	<p>Accession number: PF00585</p> <p>Definition: C-terminal domain of Threonine dehydratase</p> <p>Previous Pfam IDs: Thr_dehydratase_C;</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Bateman A</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 99.90 51.30</p> <p>Noise cutoffs: -1.10 -1.10</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98230745</p> <p>Reference Title: Structure and control of pyridoxal phosphate dependent allosteric threonine deaminase.</p> <p>Reference Author: Gallagher DT, Gilliland GL, Xiao G, Zondlo J, Fisher KE,</p> <p>Reference Author: Chinchilla D, Eisenstein E;</p> <p>Reference Location: Structure 1998;6:465-475.</p> <p>Database Reference: SCOP; 1tdj; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR001721;</p> <p>Database Reference: PDB; 1tdj ; 424; 512;</p> <p>Database Reference: PDB; 1tdj ; 329; 419;</p> <p>Comment: -!- Threonine dehydratases PALP all contain a carboxy terminal region. This region may have a regulatory role.</p> <p>Comment: Some members contain two copies of this region.</p> <p>Number of members: 30</p>
thymidylat_synt	PDOC00086	Thymidylate synthase active site	<p>Thymidylate synthase (EC 2.1.1.45) [1,2] catalyzes the reductive methylation of dUMP to dTMP with concomitant conversion of 5,10-methylenetetrahydrofolate to dihydrofolate. Thymidylate synthase plays an essential role in DNA synthesis and is an important target for certain chemotherapeutic drugs.</p> <p>Thymidylate synthase is an enzyme of about 30 to 35 Kd in most species except in protozoan and plants where it exists as a bifunctional enzyme that includes a dihydrofolate reductase domain.</p> <p>A cysteine residue is involved in the catalytic mechanism (it covalently binds the 5,6-dihydro-dUMP intermediate). The sequence around the active site of this enzyme is conserved from phages to vertebrates.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern R-x(2)-[LIVM]-x(3)-[FW]-[QN]-x(8,9)-[LV]-x-P-C-[HAVM]-x(3)-[QMT]-[FYW]-x-[LV] [C is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update</p> <p>November 1997 / Pattern and text revised.</p> <p>References</p> <p>[1]</p> <p>Benkovic S.J.</p> <p>Annu. Rev. Biochem. 49:227-251(1980).</p> <p>[2]</p> <p>Ross P., O'Gara F., Condon S.</p> <p>Appl. Environ. Microbiol. 56:2156-2163(1990).</p>
Top6A		Type II DNA topoisomerase	<p>Accession number: PF01962</p> <p>Definition: Type II DNA topoisomerase</p> <p>Author: Enright A, Ouzounis C, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Enright A</p> <p>Gathering cutoffs: -99 -99</p> <p>Trusted cutoffs: -40.40 -40.40</p>

993

			<p>Noise cutoffs: -158.40 -158.40 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97238688 Reference Title: An atypical topoisomerase II from Archaea with implications Reference Title: for meiotic recombination [see comments] Reference Author: Bergerat A, de Massy B, Gadelle D, Varoutas PC, Nicolas A, Reference Author: Forterre P; Reference Location: Nature 1997;386:414-417. Database Reference: SCOP; 1d3y; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO: IPR002815; Database Reference: PDB; 1d3y A; 77; 363; Database Reference: PDB; 1d3y B; 77; 363; Comment: Members of this family are the A subunit from type II DNA topoisomerases. Type II DNA topoisomerases catalyze the relaxation Comment: of DNA supercoiling by causing transient double strand breaks. Comment: The family includes topoisomerase VI subunit A from archaeobacteria Comment: Swiss:Q57815 EC:5.99.1.3 and SPO11 from yeast Swiss:P23179. Comment: A conserved tyrosine is thought to be involved in breaking the Comment: double stranded DNA [1]. Number of members: 9</p>
Topoisom_ba c	PDOC00333	Prokaryotic DNA topoisomeras e I active site	<p>DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type I topoisomerases act by catalyzing the transient breakage of DNA, one strand at a time, and the subsequent rejoining of the strands. When a prokaryotic type I topoisomerase breaks a DNA backbone bond, it simultaneously forms a protein-DNA link where the hydroxyl group of a tyrosine residue is joined to a 5'-phosphate on DNA, at one end of the enzyme-severed DNA strand.</p> <p>Prokaryotic organisms, such as Escherichia coli, have two type I topoisomerase isozymes: topoisomerase I (gene topA) and topoisomerase III (gene topB). Eukaryotes also contain homologs of prokaryotic topoisomerase III.</p> <p>There are a number of conserved residues in the region around the active site tyrosine; we used this region as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [EQ]-x-L-Y-[DEQST]-x(3,12)-[LIV]-[ST]-Y-x-R-[ST]-[DEQS] [The second Y is the active site tyrosine] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update December 1999 / Pattern and text revised. References [1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990). [2] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995). [3] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991). [4] Roca J. Trends Biochem. Sci. 20:156-160(1995). [E1]</p>

[illegible]

			<p>Consensus pattern [KH]-[IV]-L-[DN]-x(3)-G-x-P-A-x(2)-[IV]-x-[IV] [The K binds thyroxine]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern Y-[TH]-[IV]-[AP]-x(2)-L-S-[PQ]-[FYW]-[GS]-[FY]-[QS]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1999 / Patterns and text revised.</p> <p>References [1] Schreiber G., Richardson S.J. Comp. Biochem. Physiol. 116B:137-160(1997).</p>
TRM		N2,N2-dimethylguanosine tRNA methyltransferase	<p>Accession number: PF02005</p> <p>Definition: N2,N2-dimethylguanosine tRNA methyltransferase</p> <p>Author: Enright A, Ouzounis C, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Enright A</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 664.60 664.60</p> <p>Noise cutoffs: -259.50 -259.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmlcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98352211</p> <p>Reference Title: The tRNA(guanine-26,N2-N2) methyltransferase (Trm1) from</p> <p>Reference Title: the hyperthermophilic archaeon Pyrococcus furiosus:</p> <p>Reference Title: cloning, sequencing of the gene and its expression in</p> <p>Reference Title: Escherichia coli.</p> <p>Reference Author: Constantinesco F, Benachenhou N, Motorin Y, Grosjean H;</p> <p>Reference Location: Nucleic Acids Res 1998;26:3753-3761.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 87260951</p> <p>Reference Title: Amino-terminal extension generated from an upstream AUG</p> <p>Reference Title: codon is not required for mitochondrial import of yeast</p> <p>Reference Title: N2,N2-dimethylguanosine- specific tRNA methyltransferase.</p> <p>Reference Author: Ellis SR, Hopper AK, Martin NC;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1987;84:5172-5176.</p> <p>Database Reference: INTERPRO: IPR002905;</p> <p>Database reference: PFAMB; PB041661;</p> <p>Comment: This enzyme EC:2.1.1.32 used S-AdoMet to methylate tRNA.</p> <p>Comment: The TRM1 gene of Saccharomyces cerevisiae is necessary for</p> <p>Comment: the N2,N2-dimethylguanosine modification of both mitochondrial</p> <p>Comment: and cytoplasmic tRNAs [1]. The enzyme is found in both eukaryotes and archaeobacteria [2]</p> <p>Number of members: 10</p>
tRNA_bind		Putative tRNA binding domain	<p>Accession number: PF01588</p> <p>Definition: Putative tRNA binding domain</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_482 (release 4.1)</p> <p>Gathering cutoffs: 20 20</p> <p>Trusted cutoffs: 22.30 22.30</p> <p>Noise cutoffs: 18.20 18.20</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmlcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97306356</p> <p>Reference Title: Human tyrosyl-tRNA synthetase shares amino acid sequence</p> <p>Reference Title: homology with a putative cytokine.</p> <p>Reference Author: Kleeman TA, Wei D, Simpson KL, First EA;</p> <p>Reference Location: J Biol Chem 1997;272:14420-14425.</p>

			<p>Reference Number: [2] Reference Medline: 97050848 Reference Title: The yeast protein Arc1p binds to tRNA and functions as a cofactor for the methionyl- and glutamyl-tRNA synthetases. Reference Author: Simos G, Segref A, Fasiolo F, Hellmuth K, Shevchenko A, Reference Author: Mann M, Hurt EC; Reference Location: EMBO J 1996;15:5437-5448. Database Reference: SCOP; 1pys; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002547; Database Reference: PDB; 1b70 B; 153; 247; Database Reference: PDB; 1b7y B; 153; 247; Database Reference: PDB; 1eiy B; 153; 247; Database Reference: PDB; 1pys B; 153; 247; Database reference: PFAMB; PB010015; Comment: This domain is found in prokaryotic methionyl-tRNA synthetases, Comment: prokaryotic phenylalanyl tRNA synthetases the yeast GU4 nucleic-binding Comment: protein (G4p1 or p42, ARC1) [2], human tyrosyl-tRNA synthetase [1], Comment: and endothelial-monocyte activating polypeptide II. Comment: G4p1 binds specifically to tRNA form a complex with methionyl-tRNA Comment: synthetases [2]. In human tyrosyl-tRNA synthetase this domain may direct Comment: tRNA to the active site of the enzyme [2]. This domain may perform a Comment: common function in tRNA aminoacylation [1]. Number of members: 46</p>
tRNA-synt_2d	PDOC00363	Aminoacyl-transfer RNA synthetases class-II signatures	<p>Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure.</p> <p>The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7].</p> <p>Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. We have derived signature patterns from two of these regions.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE] Sequences known to belong to this class detected by the pattern the majority of class-II tRNA synthetases with the exception of those specific for alanine, glycine as well as bacterial histidine. Other sequence(s) detected in SWISS-PROT 43.</p> <p>Consensus pattern [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x- [LIVMSTAG]-[LIVMFY] Sequences known to belong to this class detected by the pattern the majority of class-II tRNA synthetases with the exception of those specific for serine and proline. Other sequence(s) detected in SWISS-PROT 161. Expert(s) to contact by email Cusack S. cusack@embl-grenoble.fr</p> <p>Last update July 1998 / Text revised. References [1] Schimmel P.</p>

			<p>Annu. Rev. Biochem. 56:125-158(1987).</p> <p>[2] Delarue M., Moras D. BioEssays 15:675-687(1993).</p> <p>[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).</p> <p>[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).</p> <p>[5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).</p> <p>[6] Cusack S. Biochimie 75:1077-1081(1993).</p> <p>[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).</p> <p>[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).</p>
trypsin	PDOC00124	Serine proteases, trypsin family, active sites	<p>The catalytic activity of the serine proteases from the trypsin family is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is hydrogen-bonded to a serine. The sequences in the vicinity of the active site serine and histidine residues are well conserved in this family of proteases [1]. A partial list of proteases known to belong to the trypsin family is shown below.</p> <ul style="list-style-type: none"> - Acrosin. - Blood coagulation factors VII, IX, X, XI and XII, thrombin, plasminogen, and protein C. - Cathepsin G. - Chymotrypsins. - Complement components C1r, C1s, C2, and complement factors B, D and I. - Complement-activating component of RA-reactive factor. - Cytotoxic cell proteases (granzymes A to H). - Duodenase I. - Elastases 1, 2, 3A, 3B (protease E), leukocyte (medullasin). - Enterokinase (EC 3.4.21.9) (enteropeptidase). - Hepatocyte growth factor activator. - Hepsin. - Glandular (tissue) kallikreins (including EGF-binding protein types A, B, and C, NGF-gamma chain, gamma-renin, prostate specific antigen (PSA) and tonin). - Plasma kallikrein. - Mast cell proteases (MCP) 1 (chymase) to 8. - Myeloblastin (proteinase 3) (Wegener's autoantigen). - Plasminogen activators (urokinase-type, and tissue-type). - Trypsins I, II, III, and IV. - Trypsases. - Snake venom proteases such as ancrod, batroxobin, cerastobin, flavoxobin, and protein C activator. - Collagenase from common cattle grub and collagenolytic protease from Atlantic sand fiddler crab. - Apolipoprotein(a). - Blood fluke cercarial protease. - Myeloblastin trypsin like proteases: alpha, easter, snake-locus. - Drosophila protease stubble (gene sb). - Major mite fecal allergen Der p III. <p>All the above proteins belong to family S1 in the classification of peptidases [2,E1] and originate from eukaryotic species. It should be noted that bacterial proteases that belong to family S2A are similar enough in the regions of the active site residues that they can be picked up by the same</p>

			<p>patterns. These proteases are listed below.</p> <ul style="list-style-type: none"> - <i>Achromobacter lyticus</i> protease I. - <i>Lysobacter alpha-lytic</i> protease. - Streptogrisin A and B (<i>Streptomyces</i> proteases A and B). - <i>Streptomyces griseus</i> glutamyl endopeptidase II. - <i>Streptomyces fradiae</i> proteases 1 and 2. <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-[ST]-A-[STAG]-H-C [H is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for complement components C1r and C1s, pig plasminogen, bovine protein C, rodent urokinase, ancrod, gyroxin and two insect trypsins. Other sequence(s) detected in SWISS-PROT 14.</p> <p>Consensus pattern [DNSTAGC]-[GSTAPIMVQH]-x(2)-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]-[LIVMFYSTANQH] [S is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for 18 different proteases which have lost the first conserved glycine. Other sequence(s) detected in SWISS-PROT H.influenzae protease HAP which belongs to family S6 and 3 other proteins.</p> <p>Note if a protein includes both the serine and the histidine active site signatures, the probability of it being a trypsin family serine protease is 100%</p> <p>Last update November 1997 / Text revised.</p> <p>References [1] Brenner S. Nature 334:528-530(1988).</p> <p>[2] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>
TYA		TYA transposon protein	<p>Accession number: PF01021 Definition: TYA transposon protein Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_90 (release 3.0) Gathering cutoffs: 15 15 Trusted cutoffs: 18.00 18.00 Noise cutoffs: 13.70 13.70 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97404699 Reference Title: Cryo-electron microscopy structure of yeast Ty retrotransposon virus-like particles. Reference Author: Palmer KJ, Tichelaar W, Myers N, Burns NR, Butcher SJ, Reference Author: Kingsman AJ, Fuller SD, Saibil HR; Reference Location: J Virol 1997;71:6863-6868. Database Reference INTERPRO; IPR001042; Comment: Ty are yeast transposons. A 5.7kb transcript codes for p3 a fusion protein of TYA and TYB. The TYA protein is analogous to the gag protein of retroviruses. Comment: TYA a is cleaved to form 46kd protein which can form mature virion like particles [1]. Number of members: 62</p>
tyrosinase	PDOC00398	Tyrosinase signatures	<p>Tyrosinase (EC 1.14.18.1) [1] is a copper monooxygenases that catalyzes the hydroxylation of monophenols and the oxidation of o-diphenols to o-quinols. This enzyme, found in prokaryotes as well as in eukaryotes, is involved in the formation of pigments such as melanins and other polyphenolic compounds.</p> <p>Tyrosinase binds two copper ions (CuA and CuB). Each of the two copper ion has</p>

			<p>been shown [2] to be bound by three conserved histidines residues. The regions around these copper-binding ligands are well conserved and also shared by some hemocyanins, which are copper-containing oxygen carriers from the hemolymph of many molluscs and arthropods [3,4].</p> <p>At least two proteins related to tyrosinase are known to exist in mammals:</p> <ul style="list-style-type: none"> - TRP-1 (TYRP1) [5], which is responsible for the conversion of 5,6-dihydroxyindole-2-carboxylic acid (DHICA) to indole-5,6-quinone-2-carboxylic acid. - TRP-2 (TYRP2) [6], which is the melanogenic enzyme DOPAchrome tautomerase (EC 5.3.3.12) that catalyzes the conversion of DOPAchrome to DHICA. TRP-2 differs from tyrosinases and TRP-1 in that it binds two zinc ions instead of copper [7]. <p>Other proteins that belong to this family are:</p> <ul style="list-style-type: none"> - Plants polyphenol oxidases (PPO) (EC 1.10.3.1) which catalyze the oxidation of mono- and o-diphenols to o-diquinones [8]. - <i>Caenorhabditis elegans</i> hypothetical protein C02C2.1. <p>We have derived two signature patterns for tyrosinase and related proteins. The first one contains two of the histidines that bind CuA, and is located in the N-terminal section of tyrosinase. The second pattern contains a histidine that binds CuB, that pattern is located in the central section of the enzyme.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern H-x(4,5)-F-[LIVMFTP]-x-[FW]-H-R-x(2)-[LVM]-x(3)-E [The two H's are copper ligands] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern D-P-x-F-[LIVMFYW]-x(2)-H-x(3)-D [H is a copper ligand] Sequences known to belong to this class detected by the pattern ALL the tyrosinases as well as all the hemocyanins. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update December 1999 / Patterns and text revised.</p> <p>References</p> <p>[1] Lerch K. Prog. Clin. Biol. Res. 256:85-98(1988).</p> <p>[2] Jackman M.P., Hajnal A., Lerch K. Biochem. J. 274:707-713(1991).</p> <p>[3] Linzen B. Naturwissenschaften 76:206-211(1989).</p> <p>[4] Lang W.H., van Holde K.E. Proc. Natl. Acad. Sci. U.S.A. 88:244-248(1991).</p> <p>[5] Kobayashi T., Urabe K., Winder A., Jimenez-Cervantes C., Imokawa G., Brewington T., Solano F., Garcia-Borrón J.C., Hearing V.J. EMBO J. 13:5818-5825(1994).</p> <p>[6] Jackson I.J., Chambers D.M., Tsukamoto K., Copeland N.G., Gilbert D.J., Jenkins N.A., Hearing V. EMBO J. 11:527-535(1992).</p> <p>[7] Solano F., Martinez-Liarte J.H., Jimenez-Cervantes C., Garcia-Borrón J.C., Lozano J.A.</p>
--	--	--	--

1000

			<p>Biochem. Biophys. Res. Commun. 204:1243-1250(1994).</p> <p>[8] Cary J.W., Lax A.R., Flurkey W.H. Plant Mol. Biol. 20:245-253(1992).</p>
UbiA	PDOC00727	UbiA prenyltransferase family signature	<p>The following prenyltransferases are evolutionary related [1,2]:</p> <ul style="list-style-type: none"> - Bacterial 4-hydroxybenzoate octaprenyltransferase (gene ubiA). - Yeast mitochondrial para-hydroxybenzoate--polyprenyltransferase (gene COQ2). - Protoheme IX farnesyltransferase (heme O synthase) from yeast and mammals (gene COX10) and from bacteria (genes cyoE or ctaB). <p>These proteins probably contain seven transmembrane segments. The best conserved region is located in a loop between the second and third of these segments and we used it as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern N-x(3)-[DEH]-x(2)-[LIMF]-D-x(2)-[VM]-x-R-[ST]-x(2)-R-x(4)-G</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update December 1999 / Pattern and text revised.</p> <p>References [1] Melzer M., Heide L. Biochim. Biophys. Acta 1212:93-102(1994).</p> <p>[2] Mogi T., Saiki K., Anraku Y. Mol. Microbiol. 14:391-398(1994).</p>
Ubie_methyltransferase	PDOC00911	ubiE/COQ5 methyltransferase family signatures	<p>The following methyltransferases have been shown [1] to share regions of similarities:</p> <ul style="list-style-type: none"> - Escherichia coli ubiE, which is involved in both ubiquinone and menaquinone biosynthesis and which catalyzes the S-adenosylmethionine dependent methylation of 2-polyprenyl-6-methoxy-1,4-benzoquinol into 2-polyprenyl-3-methyl-6-methoxy-1,4-benzoquinol and of demethylmenaquinol into menaquinol. - Yeast COQ5, a ubiquinone biosynthesis methyltransferase. - Bacillus subtilis spore germination protein C2 (gene: gerCB or gerC2), a probable menaquinone biosynthesis methyltransferase. - Lactococcus lactis gerC2 homolog. - Caenorhabditis elegans hypothetical protein ZK652.9. - Leishmania donovani amastigote-specific protein A41. <p>These are hydrophilic proteins of about 30 Kd (except for ZK652.9 which is 65 Kd). They can be picked up in the database by the following patterns.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern Y-D-x-M-N-x(2)-[LIVM]-S-x(3)-H-x(2)-W</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern R-V-[LIVM]-K-[PV]-[GM]-G-x-[LIVMF]-x(2)-[LIVM]-E-x-S</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update December 1999 / Pattern and text revised.</p> <p>References [1] Lee P.T., Hsu A.Y., Ha H.T., Clarke C.F. J. Bacteriol. 179:1748-1754(1997).</p>

1001

ubiquitin	PDOC00271	Ubiquitin domain signature and profile	<p>Ubiquitin [1,2,3] is a protein of seventy six amino acid residues, found in all eukaryotic cells and whose sequence is extremely well conserved from protozoan to vertebrates. It plays a key role in a variety of cellular processes, such as ATP-dependent selective degradation of cellular proteins, maintenance of chromatin structure, regulation of gene expression, stress response and ribosome biogenesis.</p> <p>In most species, there are many genes coding for ubiquitin. However they can be classified into two classes. The first class produces polyubiquitin molecules consisting of exact head to tail repeats of ubiquitin. The number of repeats is variable (up to twelve in a <i>Xenopus</i> gene). In the majority of polyubiquitin precursors, there is a final amino-acid after the last repeat. The second class of genes produces precursor proteins consisting of a single copy of ubiquitin fused to a C-terminal extension protein (CEP). There are two types of CEP proteins and both seem to be ribosomal proteins.</p> <p>Ubiquitin is a globular protein, the last four C-terminal residues (Leu-Arg-Gly-Gly) extending from the compact structure to form a 'tail', important for its function. The latter is mediated by the covalent conjugation of ubiquitin to target proteins, by an isopeptide linkage between the C-terminal glycine and the epsilon amino group of lysine residues in the target proteins.</p> <p>There are a number of proteins which are evolutionary related to ubiquitin:</p> <ul style="list-style-type: none"> - Ubiquitin-like proteins from baculoviruses as well as in some strains of bovine viral diarrhoea viruses (BVDV). These proteins are highly similar to their eukaryotic counterparts. - Mammalian protein GDX [4]. GDX is composed of two domains, a N-terminal ubiquitin-like domain of 74 residues and a C-terminal domain of 83 residues with some similarity with the thyroglobulin hormonogenic site. - Mammalian protein FAU [5]. FAU is a fusion protein which consist of a N-terminal ubiquitin-like protein of 74 residues fused to ribosomal protein S30. - Mouse protein NEDD-8 [6], a ubiquitin-like protein of 81 residues. - Human protein BAT3, a large fusion protein of 1132 residues that contains a N-terminal ubiquitin-like domain. - <i>Caenorhabditis elegans</i> protein ubl-1 [7]. Ubl-1 is a fusion protein which consist of a N-terminal ubiquitin-like protein of 70 residues fused to ribosomal protein S27A. - Yeast DNA repair protein RAD23 [8]. RAD23 contains a N-terminal domain that seems to be distantly, yet significantly, related to ubiquitin. - Mammalian RAD23-related proteins RAD23A and RAD23B. - Mammalian BCL-2 binding athanogene-1 (BAG-1). BAG-1 is a protein of 274 residues that contains a central ubiquitin-like domain. - Human spliceosome associated protein 114 (SAP 114 or SF3A120). - Yeast protein DSK2, a protein involved in spindle pole body duplication and which contains a N-terminal ubiquitin-like domain. - Human protein CKAP1/TFEB, <i>Schizosaccharomyces pombe</i> protein alp11 and <i>Caenorhabditis elegans</i> hypothetical protein F53F4.3. These proteins contain a N-terminal ubiquitin domain and a C-terminal CAP-Gly domain (see <PDOC00660>). - <i>Schizosaccharomyces pombe</i> hypothetical protein SpAC26A3.16. This protein contains a N-terminal ubiquitin domain. - Yeast protein SMT3. - Human ubiquitin-like proteins SMT3A and SMT3B. - Human ubiquitin-like protein SMT3C (also known as PIC1; Ubl1, Sumo-1; Gmp-1 or Sentrin). This protein is involved in targeting ranGAP1 to the nuclear pore complex protein ranBP2. - SMT3-like proteins in plants and <i>Caenorhabditis elegans</i>. <p>To identify ubiquitin and related proteins we have developed a pattern based on conserved positions in the central section of the sequence. A profile was also developed that spans the complete length of the ubiquitin domain.</p>
-----------	-----------	--	---

1002

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern K-x(2)-[LIVM]-x-[DESAK]-x(3)-[LIVM]-[PA]-x(3)-Q-x-[LIVM]-[LIVMC]-[LIVMFY]-x-G-x(4)-[DE]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for the RAD23 and SMT3 subfamilies, BAG-1 and SAP 114.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the profile ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.</p> <p>Last update July 1998 / Text revised.</p> <p>Bio/Technology 8:209-215(1990). References</p> <p>[1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).</p> <p>[2] Monia B.P., Ecker D.J., Croke S.T.</p> <p>[3] Finley D., Varshavsky A. Trends Biochem. Sci. 10:343-347(1985).</p> <p>[4] Filippi M., Tribioli C., Toniolo D. Genomics 7:453-457(1990).</p> <p>[5] Olvera J., Wool I.G. J. Biol. Chem. 268:17967-17974(1993).</p> <p>[6] Kumar S., Yoshida Y., Noda M. Biochem. Biophys. Res. Commun. 195:393-399(1993).</p> <p>[7] Jones D., Candido E.P. J. Biol. Chem. 268:19545-19551(1993).</p> <p>[8] Melnick L., Sherman F. J. Mol. Biol. 233:372-388(1993).</p>
UPF0004	PDOC00984	Uncharacterized protein family UPF0004 signature	<p>The following uncharacterized proteins have been shown [1] to share regions of similarities:</p> <ul style="list-style-type: none"> - Escherichia coli hypothetical protein yliG. - Escherichia coli hypothetical protein yleA and HI0019, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yqeV. - Helicobacter pylori hypothetical protein HP0269. - Helicobacter pylori hypothetical protein HP0285. - Mycoplasma iowae hypothetical protein in 16S RNA 5' region. - Mycobacterium tuberculosis hypothetical protein Rv2733c. - Rickettsia prowazekii hypothetical protein RP416. - Rickettsia prowazekii hypothetical protein RP808. - Synechocystis strain PCC 6803 hypothetical protein slr0082. - Synechocystis strain PCC 6803 hypothetical protein slI0996. - Methanococcus jannaschii hypothetical protein MJ0865. - Methanococcus jannaschii hypothetical protein MJ0867. - Caenorhabditis elegans hypothetical protein F25B5.5. <p>The size of these proteins range from 47 to 61 Kd. They contain six conserved cysteines, three of which are clustered in a region that can be used as a signature pattern.</p>

1003

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-x-[LIVMT]-x(2)-G-C-x(3)-C-[STAN]-[FY]-C-x-[LIVMT]-x(4)-G</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT 2.</p> <p>Last update December 1999 / Pattern and text revised.</p> <p>References [1] Bairoch A. Unpublished observations (1997).</p>
UPF0013		Uncharacterized membrane protein family UPF0013	<p>Accession number: PF01554</p> <p>Definition: Uncharacterized membrane protein family UPF0013</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_163 (release 4.0)</p> <p>Gathering cutoffs: -26 -26</p> <p>Trusted cutoffs: -16.10 -16.10</p> <p>Noise cutoffs: -36.70 -36.70</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Database Reference: URL; http://www.expasy.ch/cgi-bin/lists?upflist.txt;</p> <p>Database Reference INTERPRO: IPR002528;</p> <p>Database reference: PFAMB; PB041103;</p> <p>Comment: These proteins are integral membrane proteins of unknown function.</p> <p>Number of members: 47</p>
UPF0019	PDOC00949	Uncharacterized protein family UPF0019 signature	<p>The following uncharacterized proteins have been shown [1,2] to be highly similar:</p> <ul style="list-style-type: none"> - Yeast protein SNZ1, which may be involved in growth arrest and cellular response to nutrient limitation. - Yeast chromosome VI hypothetical protein YFL059w. - Yeast chromosome XIV hypothetical protein YNL333w. - Fission yeast hypothetical protein SpAC29B12.04. - Hevea brasiliensis ethylene-inducible protein HEVER. - Stellaria longipes hypothetical protein H47. - Bacillus subtilis hypothetical protein yaaD. - Haemophilus influenzae hypothetical protein HI1647. - Mycobacterium leprae hypothetical protein M1CL581.12c. - Mycobacterium tuberculosis hypothetical protein MtCY1A10.27. - Archaeoglobus fulgidus hypothetical protein AF0508. - Methanococcus jannaschii hypothetical protein MJ0677. - Methanococcus vannielii hypothetical protein in tRNA/5S rRNA gene cluster. - Methanobacterium thermoautotrophicum hypothetical protein Mth666. <p>These are hydrophilic proteins of about 32 Kd. They can be picked up in the database by the following pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern L-P-V-[VT]-[NQL]-F-[AT]-A-G-G-[LIV]-A-T-P-A-D-A-A-[LM]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1998 / Pattern and text revised.</p> <p>References [1] Sivasubramaniam S., Vanniasingham V.M., Tan C.T., Chua N.H. Plant Mol. Biol. 29:173-178(1995).</p> <p>[2] Braun E.L., Fuge E.K., Padilla P.A., Werner-Washburne M. J. Bacteriol. 178:6865-6872(1996).</p>
UPF0047	PDOC01018	Uncharacterized protein family	<p>The following uncharacterized proteins have been shown [1] to be highly similar:</p>

1004

		UPF0047 signature	<ul style="list-style-type: none"> - Bacillus subtilis hypothetical protein yugU. - Escherichia coli hypothetical protein yjbQ. - Mycobacterium tuberculosis hypothetical protein MtCY9C4.12. - Synechocystis strain PCC 6803 hypothetical protein sl1880. - Archaeoglobus fulgidus hypothetical protein AF2050. - Methanococcus jannaschii hypothetical protein MJ1081. - Methanobacterium thermoautotrophicum hypothetical protein MTH771. - Fission yeast hypothetical protein SpAC4A8.02c. <p>These are small proteins of 14 to 16 Kd. They can be picked up in the database by the following pattern. This pattern is located in the C-terminal part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern S-X(2)-[LIV]-x-[LIV]-x(2)-G-x(4)-G-T-W-Q-x-[LIV] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update July 1998 / First entry. References [1] Bairoch A. Unpublished observations (1998).</p>
UPF0052		Uncharacterized protein family UPF0052	Accession number: PF01933 Definition: Uncharacterised protein family UPF0052 Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 263.90 263.90 Noise cutoffs: -134.40 -134.40 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Database Reference INTERPRO; IPR002882; Number of members: 12
UPF0057	PDOC01013	Uncharacterized protein family UPF0057 signature	<p>The following uncharacterized proteins have been shown [1] to be evolutionary related:</p> <ul style="list-style-type: none"> - Barley low-temperature induced protein blt101. - Lophorium elongatum salt-stress induced protein ES13. - Yeast hypothetical proteins YDL123w, YDR276c, YDR525Bw and YJL151c. - Caenorhabditis elegans hypothetical proteins F47B7.1, T23F2.3, T23F2.4, T23F2.5 and ZK632.10. - Escherichia coli hypothetical protein yqaE. - Synechocystis strain PCC 6803 hypothetical protein ssr1169. <p>These are small proteins of from 52 to 140 amino-acid residues that contains two transmembrane domains. As a signature pattern we selected a region that corresponds to the end of the first transmembrane helix.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIV]-x-[STA]-[LIVF](3)-P-P-[LIVA]-[GA]-[IV]-x(4)-[GKN] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update July 1998 / First entry. References [1] Rudd K.E., Humphery-Smith I., Wasinger V.C., Bairoch A. Electrophoresis 19:536-544(1998).</p>
UPF0066	PDOC01022	Uncharacterized protein family UPF0066 signature	<p>The following uncharacterized proteins have been shown [1] to be evolutionary related:</p> <ul style="list-style-type: none"> - Escherichia coli hypothetical protein yaeB and HI0510, the corresponding Haemophilus influenzae protein.

<p>- Agrobacterium tumefaciens Ti plasmid protein virR. - Pseudomonas aeruginosa protein rcsF. - Archaeoglobus fulgidus hypothetical protein AF0241. - Archaeoglobus fulgidus hypothetical protein AF0433. - Methanococcus jannaschii hypothetical protein MJ1583. - Methanobacterium thermoautotrophicum hypothetical protein MTH1797.</p> <p>These are proteins of from 120 to 240 amino-acid residues (with the exception of AF0433 which is 366 residues long). As a signature pattern we selected a conserved region in the central part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-[AV]-F-[STA]-x-R-[SA]-x(2)-R-P-N Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update July 1999 / First entry. References [1] Bairoch A. Unpublished observations (1998).</p>	
<p>The following uncharacterized proteins have been shown [1] to share regions of similarities:</p> <ul style="list-style-type: none"> - Goat antigen UK114, a human homolog and the rat corresponding protein which is known as perchloric acid soluble protein (PSP1). PSP1 [2] may inhibit an initiation stage of cell-free protein synthesis. - Mouse heat-responsive protein HRSP12. - Yeast chromosome V hypothetical protein YER057c. - Yeast chromosome IX hypothetical protein YIL051c. - Caenorhabditis elegans hypothetical protein C23G10.2. - Escherichia coli hypothetical protein ycdK. - Escherichia coli hypothetical protein yhaR. - Escherichia coli hypothetical protein yjgF and HI0719, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yoaB. - Bacillus subtilis hypothetical protein yabJ. - Caenorhabditis elegans hypothetical protein HI1627. - Helicobacter pylori hypothetical protein HP0944. - Lactococcus lactis aldR. - Myxococcus xanthus dfrA. - Synechocystis strain PCC 6803 hypothetical protein slr0709. - Rhizobium strain NGR234 symbiotic plasmid hypothetical protein y4sK. - Pyrococcus horikoshii hypothetical protein PH0854. <p>These are small proteins of around 15 Kd whose sequence is highly conserved. As a signature pattern, we selected a well conserved region located in the C-terminal part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [PA]-[ASTPV]-R-[SACVF]-x-[LIVMFY]-x(2)-[GSAKR]-x-[LMVA]-x(5,8)-[LIVM]-E-[MI] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 4. Last update July 1999 / Pattern and text revised. References [1] Bairoch A. Unpublished observations (1995).</p>	
<p>[2] Oka T., Tsuji H., Noda C., Sakai K., Hong Y.-M., Suzuki I., Munoz S., Natori Y. J. Biol. Chem. 270:30060-30067(1995).</p>	

1006

UPF0099		Domain of unknown function UPF0099	<p>Accession number: PF01981</p> <p>Definition: Domain of unknown function UPF0099</p> <p>Previous Pfam IDs: DUF119;</p> <p>Author: Enright A, Ouzounis C, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Enright A</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 132.80 132.80</p> <p>Noise cutoffs: -35.70 -35.70</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Database Reference INTERPRO; IPR002833;</p> <p>Comment: This domain has no known function.</p> <p>Number of members: 10</p>
UQ_con	PDOC00163	Ubiquitin-conjugating enzymes active site	<p>Ubiquitin-conjugating enzymes (EC 6.3.2.19) (UBC or E2 enzymes) [1,2,3] catalyze the covalent attachment of ubiquitin to target proteins. An activated ubiquitin moiety is transferred from an ubiquitin-activating enzyme (E1) to E2 which later ligates ubiquitin directly to substrate proteins with or without the assistance of 'N-end' recognizing proteins (E3).</p> <p>In most species there are many forms of UBC (at least 9 in yeast) which are implicated in diverse cellular functions.</p> <p>A cysteine residue is required for ubiquitin-thiolester formation. There is a single conserved cysteine in UBC's and the region around that residue is conserved in the sequence of known UBC isozymes. We have used that region as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [FYWLSP]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-[LIV]-x- [LIV] [C is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for yeast UBC6 (DOA2).</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Expert(s) to contact by email Jentsch S. jentsch@zmbh.uni-heidelberg.de</p> <p>Last update July 1998 / Text revised.</p> <p>References [1] Jentsch S., Seufert W., Sommer T., Reins H.-A. Trends Biochem. Sci. 15:195-198(1990).</p> <p>[2] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).</p> <p>[3] Hershko A. Trends Biochem. Sci. 16:265-268(1991).</p>
urease_gamma	PDOC00133	Urease signatures	<p>Urease (EC 3.5.1.5) is a nickel-binding enzyme that catalyzes the hydrolysis of urea to carbon dioxide and ammonia [1]. Historically, it was the first enzyme to be crystallized (in 1926). It is mainly found in plant seeds, microorganisms and invertebrates. In plants, urease is a hexamer of identical chains. In bacteria [2], it consists of either two or three different subunits (alpha, beta and gamma).</p> <p>Urease binds two nickel ions per subunit; four histidine, an aspartate and a carbamated-lysine serve as ligands to these metals; an additional histidine is involved in the catalytic mechanism [3].</p> <p>As signatures for this enzyme, we selected a region that contains two histidine that bind one of the nickel ions and the region of the active site histidine.</p>

1007

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern T-[AY]-[GA]-[GAT]-[LIVM]-D-x-H-[LIVM]-H-x(3)-P [The two H's bind nickel] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [LIVM](2)-[CT]-H-[HN]-L-x(3)-[LIVM]-x(2)-D-[LIVM]-x-F-A [H is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References [1] Takishima K., Suga T., Mamiya G. Eur. J. Biochem. 175:151-165(1988).</p> <p>[2] Mobley H.L.T., Husinger R.P. Microbiol. Rev. 53:85-108(1989).</p> <p>[3] Jabri E., Carr M.B., Hausinger R.P., Karplus P.A. Science 268:998-1004(1995).</p>
UreD		UreD urease accessory protein	<p>Accession number: PF01774 Definition: UreD urease accessory protein Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1109 (release 4.2) Gathering cutoffs: 25 25 Trusted cutoffs: 186.00 186.00 Noise cutoffs: -42.60 -42.60 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97352660 Reference Title: Characterization of UreG, identification of a Reference Title: UreD-UreF-UreG complex, and evidence suggesting that a Reference Title: nucleotide-binding site in UreG is required for in vivo Reference Title: metallocenter assembly of Klebsiella aerogenes urease. Reference Author: Moncrief MB, Hausinger RP; Reference Location: J Bacteriol 1997;179:4081-4086. Reference Number: [2] Reference Medline: 96146510 Reference Title: Organization of Ureaplasma urealyticum urease gene cluster Reference Title: and expression in a suppressor strain of Escherichia coli. Reference Author: Neyrolles O, Ferris S, Behbahani N, Montagnier L, Blanchard Reference Author: A; Reference Location: J Bacteriol 1996;178:647-655. Reference Number: [3] Reference Medline: 94211837 Reference Title: In vitro activation of urease apoprotein and role of UreD as a chaperone required for nickel metallocenter assembly. Reference Author: Park IS, Carr MB, Hausinger RP; Reference Location: Proc Natl Acad Sci U S A 1994;91:3233-3237. Database Reference INTERPRO: IPR002669; Comment: UreD is a urease accessory protein. Urease urease hydrolyses Comment: urea into ammonia and carbamic acid [2]. UreD is involved in Comment: activation of the urease enzyme via the UreD-UreF-UreG-urease complex Comment: [1] and is required for urease nickel metallocenter assembly [3]. Comment: See also UreF UreF, UreG HypB_UreG. Number of members: 23</p>
UreF		UreF	Accession number: PF01730

1008

			<p>Definition: UreF Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_2037 (release 4.1) Gathering cutoffs: -31 -31 Trusted cutoffs: -14.30 -14.30 Noise cutoffs: -49.30 -49.30 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96404789 Reference Title: Purification and activation properties of UreD-UreF-urease Reference Title: apoprotein complexes. Reference Author: Moncrief MB, Hausinger RP; Reference Location: J Bacteriol 1996;178:5417-5421. Reference Number: [2] Reference Medline: 96146510 Reference Title: Organization of Ureaplasma urealyticum urease gene cluster Reference Title: and expression in a suppressor strain of Escherichia coli. Reference Author: Neyrolles O, Ferris S, Behbahani N, Montagnier L, Blanchard Reference Author: A; Reference Location: J Bacteriol 1996;178:647-655. Database Reference: INTERPRO; IPR002639; Comment: This family consists of the Urease accessory protein Comment: UreF. The urease enzyme (urea amidohydrolase) Comment: hydrolyses urea into ammonia and carbamic acid [2]. Comment: UreF is proposed to modulate the activation process of Comment: urease by eliminating the binding of nickel ions to Comment: noncarbamylated protein [1]. Number of members: 20</p>
XPG_N	PDOC00658	XPG protein signatures	<p>Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair. There are a minimum of seven genetic complementation groups involved in this pathway: XP-A to XP-G. The defect in XP-G can be corrected by a 133 Kd nuclear protein called XPG (or XPGC) [2].</p> <p>XPG belongs to a family of proteins [2,3,4,5,6] that are composed of two main subsets:</p> <ul style="list-style-type: none"> - Subset 1, to which belongs XPG, RAD2 from budding yeast and rad13 from fission yeast. RAD2 and XPG are single-stranded DNA endonucleases [7,8]. XPG makes the 3' incision in human DNA nucleotide excision repair [9]. - Subset 2, to which belongs mouse and human FEN-1, rad2 from fission yeast, and RAD27 from budding yeast. FEN-1 is a structure-specific endonuclease. <p>In addition to the proteins listed in the above groups, this family also includes:</p> <ul style="list-style-type: none"> - Fission yeast exo1, a 5'→3' double-stranded DNA exonuclease that could act in a pathway that corrects mismatched base pairs. - Yeast EXO1 (DHS1), a protein with probably the same function as exo1. - Yeast DIN7. <p>Sequence alignment of this family of proteins reveals that similarities are largely confined to two regions. The first is located at the N-terminal extremity (N-region) and corresponds to the first 95 to 105 amino acids. The second region is internal (I-region) and found towards the C-terminus; it spans about 140 residues and contains a highly conserved core of 27 amino acids that includes a conserved pentapeptide (E-A-[DE]-A-[QS]). It is possible that the conserved acidic residues are involved in the catalytic mechanism of DNA excision repair in XPG. The amino acids linking the N- and I-regions are not conserved; indeed, they are largely absent from proteins belonging to the second subset.</p>

			<p>We have developed two signature patterns for these proteins. The first corresponds to the central part of the N-region, the second to part of the I-region and includes the putative catalytic core pentapeptide.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [VI]-[KRE]-P-x-[FYIL]-V-F-D-G-x(2)-[PIL]-x-[LVC]-K Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [GS]-[LIVM]-[PER]-[FYS]-[LIVM]-x-A-P-x-E-A-[DE]-[PAS]-[QS]-[CLM] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Expert(s) to contact by email Clarkson S.G. clarkson@medecine.unige.ch</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References [1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).</p> <p>[2] Scherly D., Nospikel T., Corlet J., Ucla C., Bairoch A., Clarkson S.G. Nature 363:182-185(1993).</p> <p>[3] Carr A.M., Sheldrick K.S., Murray J.M., Al-Harthy R., Watts F.Z., Lehmann A.R. Nucleic Acids Res. 21:1345-1349(1993).</p> <p>[4] Murray J.M., Tavassoli M., Al-Harthy R., Sheldrick K.S., Lehmann A.R., Carr A.M., Watts F.Z. Mol. Cell. Biol. 14:4878-4888(1994).</p> <p>[5] Harrington J.J., Lieber M.R. Genes Dev. 8:1344-1355(1994).</p> <p>[6] Szankasi P., Smith G.R. Science 267:1166-1169(1995).</p> <p>[7] Habraken Y., Sung P., Prakash L., Prakash S. Nature 366:365-368(1993).</p> <p>[8] O'Donovan A., Scherly D., Clarkson S.G., Wood R.D. J. Biol. Chem. 269:15965-15968(1994).</p> <p>[9] O'Donovan A., Davies A.A., Moggs J.G., West S.C., Wood R.D. Nature 371:432-435(1994).</p>
Y_phosphatase	PDOC00323	Tyrosine specific protein phosphatases signature and profiles	<p>Tyrosine specific protein phosphatases (EC 3.1.3.48) (PTPase) [1 to 5] are enzymes that catalyze the removal of a phosphate group attached to a tyrosine residue. These enzymes are very important in the control of cell growth, proliferation, differentiation and transformation. Multiple forms of PTPase have been characterized and can be classified into two categories: soluble PTPases and transmembrane receptor proteins that contain PTPase domain(s). The currently known PTPases are listed below:</p> <p>Soluble PTPases.</p> <ul style="list-style-type: none"> - PTPN1 (PTP-1B). - PTPN2 (T-cell PTPase; TC-PTP). - PTPN3 (H1) and PTPN4 (MEG), enzymes that contain an N-terminal band

4.1- like domain (see <PDOC00566>) and could act at junctions between the membrane and cytoskeleton.

- PTPN5 (STEP).
- PTPN6 (PTP-1C; HCP; SHP) and PTPN11 (PTP-2C; SH-PTP3; Syp), enzymes which contain two copies of the SH2 domain at its N-terminal extremity. The Drosophila protein corkscrew (gene csw) also belongs to this subgroup.
- PTPN7 (LC-PTP; Hematopoietic protein-tyrosine phosphatase; HePTP).
- PTPN8 (70Z-PEP).
- PTPN9 (MEG2).
- PTPN12 (PTP-G1; PTP-P19).
- Yeast PTP1.
- Yeast PTP2 which may be involved in the ubiquitin-mediated protein degradation pathway.
- Fission yeast pyp1 and pyp2 which play a role in inhibiting the onset of mitosis.
- Fission yeast pyp3 which contributes to the dephosphorylation of cdc2.
- Yeast CDC14 which may be involved in chromosome segregation.
- Yersinia virulence plasmid PTPases (gene yopH).
- Autographa californica nuclear polyhedrosis virus 19 Kd PTPase.

Dual specificity PTPases.

- DUSP1 (PTPN10; MAP kinase phosphatase-1; MKP-1); which dephosphorylates MAP kinase on both Thr-183 and Tyr-185.
- DUSP2 (PAC-1), a nuclear enzyme that dephosphorylates MAP kinases ERK1 and ERK2 on both Thr and Tyr residues.
- DUSP3 (VHR).
- DUSP4 (HVH2).
- DUSP5 (HVH3).
- DUSP6 (Pyst1; MKP-3).
- DUSP7 (Pyst2; MKP-X).
- Yeast MSG5, a PTPase that dephosphorylates MAP kinase FUS3.
- Yeast YVH1.
- Vaccinia virus H1 PTPase; a dual specificity phosphatase.

Receptor PTPases.

Structurally, all known receptor PTPases, are made up of a variable length extracellular domain, followed by a transmembrane region and a C-terminal catalytic cytoplasmic domain. Some of the receptor PTPases contain fibronectin type III (FN-III) repeats, immunoglobulin-like domains, MAM domains or carbonic anhydrase-like domains in their extracellular region. The cytoplasmic region generally contains two copies of the PTPase domain. The first seems to have enzymatic activity, while the second is inactive but seems to affect substrate specificity of the first. In these domains, the catalytic cysteine is generally conserved but some other, presumably important, residues are not.

In the following table, the domain structure of known receptor PTPases is shown:

	Extracellular		Intracellular			
	Ig	FN-3	CAH	MAM	PTPase	
Leukocyte common antigen (LCA) (CD45)			0	2	0	2
Leukocyte antigen related (LAR)			3	8	0	2
Drosophila DLAR			3	9	0	2
Drosophila DPTP			2	2	0	2
PTP-alpha (LRP)			0	0	0	2
PTP-beta			0	16	0	1
PTP-gamma			0	1	1	2
PTP-delta			0	>7	0	2
PTP-epsilon			0	0	0	2
PTP-kappa			1	4	0	2
PTP-mu			1	4	0	2
PTP-zeta			0	1	1	2

PTPase domains consist of about 300 amino acids. There are two conserved

1011

			<p>cysteines, the second one has been shown to be absolutely required for activity. Furthermore, a number of conserved residues in its immediate vicinity have also been shown to be important.</p> <p>We derived a signature pattern for PTPase domains centered on the active site cysteine.</p> <p>There are three profiles for PTPases, the first one spans the complete domain and is not specific to any subtype. The second profile is specific to dual-specificity PTPases and the third one to the PTP subfamily.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMF]-H-C-x(2)-G-x(3)-[STC]-[STAGP]-x-[LIVMFY] [C is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for nine sequences.</p> <p>Other sequence(s) detected in SWISS-PROT 3.</p> <p>Sequences known to belong to this class detected by the 1st profile ALL.</p> <p>Other sequence(s) detected in SWISS-PROT 2.</p> <p>Sequences known to belong to this class detected by the 2nd profile ALL dual type PTPases.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the 3rd profile ALL PTP type PTPases.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note the M-phase inducer phosphatases (cdc25-type phosphatase) are tyrosine- protein phosphatases that are not structurally related to the above PTPases.</p> <p>Note this documentation entry is linked to both a signature pattern and to profiles. As profiles are much more sensitive than the pattern, you should use them if you have access to the necessary software tools to do so.</p> <p>Last update July 1999 / Text revised.</p> <p>References</p> <p>[1] Fischer E.H., Charbonneau H., Tonks N.K. Science 253:401-406(1991).</p> <p>[2] Charbonneau H., Tonks N.K. Annu. Rev. Cell Biol. 8:463-493(1992).</p> <p>[3] Trowbridge I.S. J. Biol. Chem. 266:23517-23520(1991).</p> <p>[4] Tonks N.K., Charbonneau H. Trends Biochem. Sci. 14:497-500(1989).</p> <p>[5] Hunter T. Cell 58:1013-1016(1989).</p>
Zein		Zein seed storage protein	<p>Accession number: PF01559</p> <p>Definition: Zein seed storage protein</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_181 (release 4.0)</p> <p>Gathering cutoffs: -21 -21</p> <p>Trusted cutoffs: 4.60 4.60</p> <p>Noise cutoffs: -46.60 -46.60</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 93197294</p>

1012

			<p>Reference Title: Studies of the zein-like alpha-prolamins based on an analysis of amino acid sequences: implications for their evolution and three-dimensional structure.</p> <p>Reference Author: Garratt R, Oliva G, Caracelli I, Leite A, Arruda P;</p> <p>Reference Location: Proteins 1993;15:88-99.</p> <p>Database Reference: INTERPRO; IPR002530;</p> <p>Comment: Zeins are seed storage proteins. They are unusually rich in</p> <p>Comment: glutamine, proline, alanine, and leucine residues and their sequences show a series of tandem repeats [1].</p> <p>Number of members: 48</p>
zf-AN1		AN1-like Zinc finger	<p>Accession number: PF01428</p> <p>Definition: AN1-like Zinc finger</p> <p>Author: Bateman A, SMART</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: SMART</p> <p>Gathering cutoffs: 16 16</p> <p>Trusted cutoffs: 16.40 16.40</p> <p>Noise cutoffs: 7.30 7.30</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 93292985</p> <p>Reference Title: Two related localized mRNAs from <i>Xenopus laevis</i> encode</p> <p>Reference Title: ubiquitin-like fusion proteins.</p> <p>Reference Author: Linnen JM, Bailey CP, Weeks DL;</p> <p>Reference Location: Gene 1993;128:181-188.</p> <p>Database reference: SMART; ZnF_AN1;</p> <p>Database Reference: INTERPRO; IPR000058;</p> <p>Comment: Zinc finger at the C-terminus of An1 Swiss:Q91889, a ubiquitin-like</p> <p>Comment: protein in <i>Xenopus laevis</i>.</p> <p>Comment: The following pattern describes the zinc finger.</p> <p>Comment: C-X2-C-X(9-12)-C-X(1-2)-C-X4-C-X2-H-X5-H-X-C</p> <p>Comment: Where X can be any amino acid, and numbers in brackets indicate the number of residues.</p> <p>Number of members: 18</p>
zf-CONSTANS		CONSTANS family zinc finger	<p>Accession number: PF01760</p> <p>Definition: CONSTANS family zinc finger</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1072 (release 4.2)</p> <p>Gathering cutoffs: 25 10</p> <p>Trusted cutoffs: 76.10 17.20</p> <p>Noise cutoffs: 9.70 9.70</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95211836</p> <p>Reference Title: The CONSTANS gene of <i>Arabidopsis</i> promotes flowering and</p> <p>Reference Title: encodes a protein showing similarities to zinc finger</p> <p>Reference Title: transcription factors.</p> <p>Reference Author: Putterill J, Robson F, Lee K, Simon R, Coupland G;</p> <p>Reference Location: Cell 1995;80:847-857.</p> <p>Database Reference: INTERPRO; IPR002926;</p> <p>Number of members: 45</p>
zf-DHHC		DHHC zinc finger domain	<p>Accession number: PF01529</p> <p>Definition: DHHC zinc finger domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_945 (release 4.0)</p> <p>Gathering cutoffs: 22 22</p> <p>Trusted cutoffs: 22.40 22.40</p> <p>Noise cutoffs: -22.40 -22.40</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 99250263</p> <p>Reference Title: The drosophila STAM gene homolog is in a tight gene</p>

1013

			<p>Reference Title: cluster, and its expression correlates to that of the adjacent gene ial.</p> <p>Reference Author: Mesilaty-Gross S, Reich A, Motro B, Wides R;</p> <p>Reference Location: Gene 1999;231:173-186.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97315340</p> <p>Reference Title: Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins.</p> <p>Reference Author: Bohm S, Frishman D, Mewes HW;</p> <p>Reference Location: Nucleic Acids Res 1997;25:2464-2469.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 99321009</p> <p>Reference Title: The DHHC domain: a new highly conserved cysteine-rich motif.</p> <p>Reference Author: Putilina T, Wong P, Gentleman S;</p> <p>Reference Location: Mol Cell Biochem 1999;195:219-226.</p> <p>Reference Number: [4]</p> <p>Reference Medline: 10490616</p> <p>Reference Title: Erf2, a Novel Gene Product That Affects the Localization and Palmitoylation of Ras2 in Saccharomyces cerevisiae.</p> <p>Reference Author: Bartels DJ, Mitchell DA, Dong X, Deschenes RJ;</p> <p>Reference Location: Mol Cell Biol 1999;19:6775-6787.</p> <p>Database Reference: INTERPRO; IPR001594;</p> <p>Comment: This domain is also known as NEW1 [2]. This domain is predicted to be a zinc binding domain. The function of this domain is unknown, but it has been predicted to be involved in protein-protein or protein-DNA interactions [3].</p> <p>Number of members: 34</p>
zf-MYND		MYND finger	<p>Accession number: PF01753</p> <p>Definition: MYND finger</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: Bateman A</p> <p>Gathering cutoffs: 11 11</p> <p>Trusted cutoffs: 17.30 17.30</p> <p>Noise cutoffs: 5.50 5.50</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96203118</p> <p>Reference Title: DEAF-1, a novel protein that binds an essential region in a</p> <p>Reference Title: Deformed response element.</p> <p>Reference Author: Gross CT, McGinnis W;</p> <p>Reference Location: EMBO J 1996;15:1961-1970.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98079069</p> <p>Reference Title: Molecular cloning, sequence analysis, expression, and tissue distribution of suppressin, a novel suppressor of cell cycle entry.</p> <p>Reference Author: LeBoeuf RD, Ban EM, Green MM, Stone AS, Propst SM, Blalock</p> <p>Reference Author: JE, Tauber JD;</p> <p>Reference Location: J Biol Chem 1998;273:361-368.</p> <p>Database Reference: INTERPRO; IPR002893;</p> <p>Number of members: 48</p>
Zn_carbOpept	PDOC00123	Zinc carboxypeptidases, zinc-binding regions signatures	<p>There are a number of different types of zinc-dependent carboxypeptidases (EC 3.4.17.-) [1,2]. All these enzymes seem to be structurally and functionally related. The enzymes that belong to this family are listed below.</p> <ul style="list-style-type: none"> - Carboxypeptidase A1 (EC 3.4.17.1), a pancreatic digestive enzyme that can removes all C-terminal amino acids with the exception of Arg, Lys and Pro. - Carboxypeptidase A2 (EC 3.4.17.15), a pancreatic digestive enzyme with a specificity similar to that of carboxypeptidase A1, but with a preference for bulkier C-terminal residues. - Carboxypeptidase B (EC 3.4.17.2), also a pancreatic digestive enzyme, but that preferentially removes C-terminal Arg and Lys. - Carboxypeptidase N (EC 3.4.17.3) (also known as arginine carboxypeptidase), a plasma enzyme which protects the body from potent vasoactive and

1014

			<p>inflammatory peptides containing C-terminal Arg or Lys (such as kinins or anaphylatoxins) which are released into the circulation.</p> <ul style="list-style-type: none"> - Carboxypeptidase H (EC 3.4.17.10) (also known as enkephalin convertase or carboxypeptidase E), an enzyme located in secretory granules of pancreatic islets, adrenal gland, pituitary and brain. This enzyme removes residual C-terminal Arg or Lys remaining after initial endoprotease cleavage during prohormone processing. - Carboxypeptidase M (EC 3.4.17.12), a membrane bound Arg and Lys specific enzyme. <p>It is ideally situated to act on peptide hormones at local tissue sites where it could control their activity before or after interaction with specific plasma membrane receptors.</p> <ul style="list-style-type: none"> - Mast cell carboxypeptidase (EC 3.4.17.1), an enzyme with a specificity to carboxypeptidase A, but found in the secretory granules of mast cells. - <i>Streptomyces griseus</i> carboxypeptidase (Cpase SG) (EC 3.4.17.-) [3], which combines the specificities of mammalian carboxypeptidases A and B. - <i>Thermoactinomyces vulgaris</i> carboxypeptidase T (EC 3.4.17.18) (CPT) [4], which also combines the specificities of carboxypeptidases A and B. - AEBP1 [5], a transcriptional repressor active in preadipocytes. AEBP1 seems to regulate transcription by cleavage of other transcriptional proteins. - Yeast hypothetical protein YHR132c. <p>All of these enzymes bind an atom of zinc. Three conserved residues are implicated in the binding of the zinc atom: two histidines and a glutamic acid. We have derived two signature patterns which contain these three zinc-ligands.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [PK]-x-[LIVMFY]-x-[LIVMFY]-x(4)-H-[STAG]-x-E-x-[LIVM]-[STAG]-x(6)-[LIVMFYTA] [H and E are zinc ligands] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT <i>Bacillus sphaericus</i> endopeptidase I which hydrolyses the gamma-D-Glu-(L)meso-diaminopimelic acid bond of spore cortex peptidoglycan [6] and which is possibly distantly related to zinc carboxypeptidases.</p> <p>Consensus pattern H-[STAG]-x(3)-[LIVME]-x(2)-[LIVMFYW]-P-[FYW] [H is a zinc ligand] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 40.</p> <p>Note if a protein includes both signatures, the probability of it being a eukaryotic zinc carboxypeptidase is 100%</p> <p>Note these proteins belong to families M14A/M14B in the classification of peptidases [7,E1]. Last update November 1995 / Patterns and text revised. References [1] Tan F., Chan S.J., Steiner D.F., Schilling J.W., Skidgel R.A. J. Biol. Chem. 264:13165-13170(1989). [2] Reynolds D.S., Stevens R.L., Gurley D.S., Lane W.S., Austen K.F., Serafin W.E. J. Biol. Chem. 264:20094-20099(1989). [3] Narahashi Y. J. Biochem. 107:879-886(1990). [4] Teplyakov A., Polyakov K., Obmolova G., Strokopytov B., Kuranova I., Osterman A.L., Grishin N.V., Smulevitch S.V., Zagnitko O.P., Galperina O.V., Matz M.V., Stepanov V.M. Eur. J. Biochem. 208:281-288(1992). [5] He G.-P., Muise A., Li A.W., Ro H.-S.</p>
--	--	--	---

581

			<p>Nature 378:92-96(1995).</p> <p>[6] Hourdou M.-L., Guinand M., Vacheron M.J., Michel G., Denoroy L., Duez C.M., Englebert S., Joris B., Weber G., Ghuysen J.-M. Biochem. J. 292:563-570(1993).</p> <p>[7] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>
ZZ		Zinc finger present in dystrophin, CBP/p300	<p>Accession number: PF00569</p> <p>Definition: Zinc finger present in dystrophin, CBP/p300</p> <p>Author: SMART</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: Alignment kindly provided by SMART</p> <p>Gathering cutoffs: 14 14</p> <p>Trusted cutoffs: 14.60 14.60</p> <p>Noise cutoffs: 10.90 10.90</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96402609</p> <p>Reference Title: ZZ and TAZ: new putative zinc fingers in dystrophin and other proteins.</p> <p>Reference Author: Ponting CP, Blake DJ, Davies KE, Kendrick-Jones J, Winder</p> <p>Reference Author: SJ;</p> <p>Reference Location: Trends Biochem Sci 1996;21:11-13.</p> <p>Database Reference: EXPERT; Chris.Ponting@human-anatomy.oxford.ac.uk;</p> <p>Database Reference: INTERPRO; IPR000433;</p> <p>Database reference: PFAMB; PB041629;</p> <p>Comment: ZZ in dystrophin binds calmodulin</p> <p>Comment: Putative zinc finger; binding not yet shown.</p> <p>Number of members: 87</p>

AA. Activities of Polypeptides Comprising Signal Peptides

Polypeptides comprising signal peptides are a family of proteins that are typically
5 targeted to (1) a particular organelle or intracellular compartment, (2) interact with a
particular molecule or (3) for secretion outside of a host cell. Example of polypeptides
comprising signal peptides include, without limitation, secreted proteins, soluble proteins,
receptors, proteins retained in the ER, etc.

10 These proteins comprising signal peptides are useful to modulate ligand-receptor
interactions, cell-to-cell communication, signal transduction, intracellular communication,
and activities and/or chemical cascades that take part in an organism outside or within of any
particular cell.

15 One class of such proteins are soluble proteins which are transported out of the cell.
These proteins can act as ligands that bind to receptor to trigger signal transduction or to
permit communication between cells.

20 Another class is receptor proteins which also comprise a retention domain that lodges
the receptor protein in the membrane when the cell transports the receptor to the surface of
the cell. Like the soluble ligands, receptors can also modulate signal transduction and
communication between cells.

25 In addition the signal peptide itself can serve as a ligand for some receptors. An
example is the interaction of the ER targeting signal peptide with the signal recognition
particle (SRP). Here, the SRP binds to the signal peptide, halting translation, and the
resulting SRP complex then binds to docking proteins located on the surface of the ER,
prompting transfer of the protein into the ER.

30 A description of signal peptide residue composition is described below in Subsection
IV.C.1.

III. Methods of Modulating Polypeptide Production

It is contemplated that polynucleotides of the invention can be incorporated into a host cell or in-vitro system to modulate polypeptide production. For instance, the SDFs prepared as described herein can be used to prepare expression cassettes useful in a number of techniques for suppressing or enhancing expression.

An example are polynucleotides comprising sequences to be transcribed, such as coding sequences, of the present invention can be inserted into nucleic acid constructs to modulate polypeptide production. Typically, such sequences to be transcribed are heterologous to at least one element of the nucleic acid construct to generate a chimeric gene or construct.

Another example of useful polynucleotides are nucleic acid molecules comprising regulatory sequences of the present invention. Chimeric genes or constructs can be generated when the regulatory sequences of the invention linked to heterologous sequences in a vector construct. Within the scope of invention are such chimeric gene and/or constructs.

Also within the scope of the invention are nucleic acid molecules, whereof at least a part or fragment of these DNA molecules are presented in Tables 1 and 2 of the present application, and wherein the coding sequence is under the control of its own promoter and/or its own regulatory elements. Such molecules are useful for transforming the genome of a host cell or an organism regenerated from said host cell for modulating polypeptide production.

Additionally, a vector capable of producing the oligonucleotide can be inserted into the host cell to deliver the oligonucleotide.

More detailed description of components to be included in vector constructs are described both above and below.

Whether the chimeric vectors or native nucleic acids are utilized, such polynucleotides can be incorporated into a host cell to modulate polypeptide production. Native genes and/or nucleic acid molecules can be effective when exogenous to the host cell.

Methods of modulating polypeptide expression includes, without limitation:

Suppression methods, such as

Antisense

Ribozymes

Co-suppression

Insertion of Sequences into the Gene to be Modulated

Regulatory Sequence Modulation.

as well as Methods for Enhancing Production, such as
Insertion of Exogenous Sequences; and
Regulatory Sequence Modulation.

III.A. Suppression

Expression cassettes of the invention can be used to suppress expression of endogenous genes which comprise the SDF sequence. Inhibiting expression can be useful, for instance, to tailor the ripening characteristics of a fruit (Oeller et al., *Science* 254:437 (1991)) or to influence seed size (WO98/07842) or to provoke cell ablation (Mariani et al., *Nature* 357: 384-387 (1992)).

As described above, a number of methods can be used to inhibit gene expression in plants, such as antisense, ribozyme, introduction of exogenous genes into a host cell, insertion of a polynucleotide sequence into the coding sequence and/or the promoter of the endogenous gene of interest, and the like.

III.A.1. Antisense

An expression cassette as described above can be transformed into host cell or plant to produce an antisense strand of RNA. For plant cells, antisense RNA inhibits gene expression by preventing the accumulation of mRNA which encodes the enzyme of interest, *see*, e.g., Sheehy et al., *Proc. Nat. Acad. Sci. USA*, 85:8805 (1988), and Hiatt et al., U.S. Patent No. 4,801,340.

III.A.2. Ribozymes

Similarly, ribozyme constructs can be transformed into a plant to cleave mRNA and down-regulate translation.

III.A.3. Co-Suppression

Another method of suppression is by introducing an exogenous copy of the gene to be suppressed. Introduction of expression cassettes in which a nucleic acid is configured in the sense orientation with respect to the promoter has been shown to prevent the accumulation of mRNA. A detailed description of this method is described above.

III.A.4. Insertion of Sequences into the Gene to be Modulated

Yet another means of suppressing gene expression is to insert a polynucleotide into the gene of interest to disrupt transcription or translation of the gene.

Homologous recombination could be used to target a polynucleotide insert to a gene using the Cre-Lox system (A.C. Vergunst et al., *Nucleic Acids Res.* 26:2729 (1998), A.C. Vergunst et al., *Plant Mol. Biol.* 38:393 (1998), H. Albert et al., *Plant J.* 7:649 (1995)).

In addition, random insertion of polynucleotides into a host cell genome can also be used to disrupt the gene of interest. Azpiroz-Leehan et al., *Trends in Genetics* 13:152 (1997). In this method, screening for clones from a library containing random insertions is preferred for identifying those that have polynucleotides inserted into the gene of interest. Such screening can be performed using probes and/or primers described above based on sequences from Tables 1 and 2, fragments thereof, and substantially similar sequence thereto. The screening can also be performed by selecting clones or any transgenic plants having a desired phenotype.

III.A.5. Regulatory Sequence Modulation

The SDFs described in Tables 1 and 2, and fragments thereof are examples of nucleotides of the invention that contain regulatory sequences that can be used to suppress or inactivate transcription and/or translation from a gene of interest as discussed in I.C.5.

III.A.6. Genes Comprising Dominant-Negative Mutations

When suppression of production of the endogenous, native protein is desired it is often helpful to express a gene comprising a dominant negative mutation. Production of protein variants produced from genes comprising dominant negative mutations is a useful tool for research. Genes comprising dominant negative mutations can produce a variant polypeptide which is capable of competing with the native polypeptide, but which does not produce the native result. Consequently, over expression of genes comprising these mutations can titrate out an undesired activity of the native protein. For example, The product from a gene comprising a dominant negative mutation of a receptor can be used to constitutively activate or suppress a signal transduction cascade, allowing examination of the phenotype and thus the trait(s) controlled by that receptor and pathway. Alternatively, the protein arising from the gene comprising a dominant-negative mutation can be an inactive enzyme still capable

of binding to the same substrate as the native protein and therefore competes with such native protein.

Products from genes comprising dominant-negative mutations can also act upon the native protein itself to prevent activity. For example, the native protein may be active only as a homo-multimer or as one subunit of a hetero-multimer. Incorporation of an inactive subunit into the multimer with native subunit(s) can inhibit activity.

Thus, gene function can be modulated in host cells of interest by insertion into these cells vector constructs comprising a gene comprising a dominant-negative mutation.

III.B. Enhanced Expression

Enhanced expression of a gene of interest in a host cell can be accomplished by either (1) insertion of an exogenous gene; or (2) promoter modulation.

III.B.1. Insertion of an Exogenous Gene

Insertion of an expression construct encoding an exogenous gene can boost the number of gene copies expressed in a host cell.

Such expression constructs can comprise genes that either encode the native protein that is of interest or that encode a variant that exhibits enhanced activity as compared to the native protein. Such genes encoding proteins of interest can be constructed from the sequences from Tables 1 and 2, fragments thereof, and substantially similar sequence thereto.

Such an exogenous gene can include either a constitutive promoter permitting expression in any cell in a host organism or a promoter that directs transcription only in particular cells or times during a host cell life cycle or in response to environmental stimuli.

III.B.2. Regulatory Sequence Modulation

The SDFs of Tables 1 and 2, and fragments thereof, contain regulatory sequences that can be used to enhance expression of a gene of interest. For example, some of these sequences contain useful enhancer elements. In some cases, duplication of enhancer elements or insertion of exogenous enhancer elements will increase expression of a desired gene from a particular promoter. As other examples, all promoters require binding of a regulatory protein to be activated, while some promoters may need a protein that signals a promoter binding protein to expose a polymerase binding site. In either case, over-production of such proteins

can be used to enhance expression of a gene of interest by increasing the activation time of the promoter.

Such regulatory proteins are encoded by some of the sequences in Tables 1 and 2, fragments thereof, and substantially similar sequences thereto.

Coding sequences for these proteins can be constructed as described above.

IV. Gene Constructs and Vector Construction

To use isolated SDFs of the present invention or a combination of them or parts and/or mutants and/or fusions of said SDFs in the above techniques, recombinant DNA vectors which comprise said SDFs and are suitable for transformation of cells, such as plant cells, are usually prepared. The SDF construct can be made using standard recombinant DNA techniques (Sambrook et al. 1989) and can be introduced to the species of interest by *Agrobacterium*-mediated transformation or by other means of transformation (e.g., particle gun bombardment) as referenced below.

The vector backbone can be any of those typical in the art such as plasmids, viruses, artificial chromosomes, BACs, YACs and PACs and vectors of the sort described by

- (a) **BAC:** Shizuya et al., Proc. Natl. Acad. Sci. USA 89: 8794-8797 (1992); Hamilton et al., Proc. Natl. Acad. Sci. USA 93: 9975-9979 (1996);
- (b) **YAC:** Burke et al., Science 236:806-812 (1987);
- (c) **PAC:** Sternberg N. et al., Proc Natl Acad Sci U S A. Jan;87(1):103-7 (1990);
- (d) **Bacteria-Yeast Shuttle Vectors:** Bradshaw et al., Nucl Acids Res 23: 4850-4856 (1995);
- (e) **Lambda Phage Vectors:** Replacement Vector, e.g., Frischauf et al., J. Mol Biol 170: 827-842 (1983); or Insertion vector, e.g., Huynh et al., In: Glover NM (ed) DNA Cloning: A practical Approach, Vol.1 Oxford: IRL Press (1985);
- (f) **T-DNA gene fusion vectors :**Walden et al., Mol Cell Biol 1: 175-194 (1990); and
- (g) **Plasmid vectors:** Sambrook et al., infra.

Typically, a vector will comprise the exogenous gene, which in its turn comprises an SDF of the present invention to be introduced into the genome of a host cell, and which gene may be an antisense construct, a ribozyme construct chimera, or a coding sequence with

any desired transcriptional and/or translational regulatory sequences, such as promoters, UTRs, and 3' end termination sequences. Vectors of the invention can also include origins of replication, scaffold attachment regions (SARs), markers, homologous sequences, introns, etc.

A DNA sequence coding for the desired polypeptide, for example a cDNA sequence encoding a full length protein, will preferably be combined with transcriptional and translational initiation regulatory sequences which will direct the transcription of the sequence from the gene in the intended tissues of the transformed plant.

For example, for over-expression, a plant promoter fragment may be employed that will direct transcription of the gene in all tissues of a regenerated plant. Alternatively, the plant promoter may direct transcription of an SDF of the invention in a specific tissue (tissue-specific promoters) or may be otherwise under more precise environmental control (inducible promoters).

If proper polypeptide production is desired, a polyadenylation region at the 3'-end of the coding region is typically included. The polyadenylation region can be derived from the natural gene, from a variety of other plant genes, or from T-DNA.

The vector comprising the sequences from genes or SDF or the invention may comprise a marker gene that confers a selectable phenotype on plant cells. The vector can include promoter and coding sequence, for instance. For example, the marker may encode biocide resistance, particularly antibiotic resistance, such as resistance to kanamycin, G418, bleomycin, hygromycin, or herbicide resistance, such as resistance to chlorosulfuron or phosphinotricin.

IV.A. Coding Sequences

Generally, the sequence in the transformation vector and to be introduced into the genome of the host cell does not need to be absolutely identical to an SDF of the present invention. Also, it is not necessary for it to be full length, relative to either the primary transcription product or fully processed mRNA. Furthermore, the introduced sequence need not have the same intron or exon pattern as a native gene. Also, heterologous non-coding segments can be incorporated into the coding sequence without changing the desired amino acid sequence of the polypeptide to be produced.

IV.B. Promoters

As explained above, introducing an exogenous SDF from the same species or an orthologous SDF from another species can modulate the expression of a native gene corresponding to that SDF of interest. Such an SDF construct can be under the control of either a constitutive promoter or a highly regulated inducible promoter (*e.g.*, a copper inducible promoter). The promoter of interest can initially be either endogenous or heterologous to the species in question. When re-introduced into the genome of said species, such promoter becomes exogenous to said species. Over-expression of an SDF transgene can lead to co-suppression of the homologous endogeneous sequence thereby creating some alterations in the phenotypes of the transformed species as demonstrated by similar analysis of the chalcone synthase gene (Napoli et al., *Plant Cell* 2:279 (1990) and van der Krol et al., *Plant Cell* 2:291 (1990)). If an SDF is found to encode a protein with desirable characteristics, its over-production can be controlled so that its accumulation can be manipulated in an organ- or tissue-specific manner utilizing a promoter having such specificity.

Likewise, if the promoter of an SDF (or an SDF that includes a promoter) is found to be tissue-specific or developmentally regulated, such a promoter can be utilized to drive or facilitate the transcription of a specific gene of interest (*e.g.*, seed storage protein or root-specific protein). Thus, the level of accumulation of a particular protein can be manipulated or its spatial localization in an organ- or tissue- specific manner can be altered.

IV. C Signal Peptides

SDFs of the present invention containing signal peptides are indicated in Tables 1 and 2. In some cases it may be desirable for the protein encoded by an introduced exogenous or orthologous SDF to be targeted (1) to a particular organelle intracellular compartment, (2) to interact with a particular molecule such as a membrane molecule or (3) for secretion outside of the cell harboring the introduced SDF. This will be accomplished using a signal peptide.

Signal peptides direct protein targeting, are involved in ligand-receptor interactions and act in cell to cell communication. Many proteins, especially soluble proteins, contain a signal peptide that targets the protein to one of several different intracellular compartments. In plants, these compartments include, but are not limited to, the endoplasmic reticulum (ER), mitochondria, plastids (such as chloroplasts), the vacuole, the Golgi apparatus, protein storage vessicles (PSV) and, in general, membranes. Some signal peptide sequences are conserved, such as the Asn-Pro-Ile-Arg amino acid motif found in the N-terminal propeptide

signal that targets proteins to the vacuole (Marty (1999) *The Plant Cell* 11: 587-599). Other signal peptides do not have a consensus sequence *per se*, but are largely composed of hydrophobic amino acids, such as those signal peptides targeting proteins to the ER (Vitale and Denecke (1999) *The Plant Cell* 11: 615-628). Still others do not appear to contain either

5 a consensus sequence or an identified common secondary sequence, for instance the chloroplast stromal targeting signal peptides (Keegstra and Cline (1999) *The Plant Cell* 11: 557-570). Furthermore, some targeting peptides are bipartite, directing proteins first to an organelle and then to a membrane within the organelle (e.g. within the thylakoid lumen of the chloroplast; see Keegstra and Cline (1999) *The Plant Cell* 11: 557-570). In addition to the

10 diversity in sequence and secondary structure, placement of the signal peptide is also varied. Proteins destined for the vacuole, for example, have targeting signal peptides found at the N-terminus, at the C-terminus and at a surface location in mature, folded proteins. Signal peptides also serve as ligands for some receptors.

These characteristics of signal proteins can be used to more tightly control the phenotypic expression of introduced SDFs. In particular, associating the appropriate signal sequence with a specific SDF can allow sequestering of the protein in specific organelles (plastids, as an example), secretion outside of the cell, targeting interaction with particular receptors, etc. Hence, the inclusion of signal proteins in constructs involving the SDFs of the invention increases the range of manipulation of SDF phenotypic expression. The nucleotide

15 sequence of the signal peptide can be isolated from characterized genes using common molecular biological techniques or can be synthesized *in vitro*.

In addition, the native signal peptide sequences, both amino acid and nucleotide, described in Tables 1 and 2 can be used to modulate polypeptide transport. Further variants of the native signal peptides described in Tables 1 and 2 are contemplated. Insertions, deletions, or substitutions can be made. Such variants will retain at least one of the functions of the native signal peptide as well as exhibiting some degree of sequence identity to the native sequence.

Also, fragments of the signal peptides of the invention are useful and can be fused with other signal peptides of interest to modulate transport of a polypeptide.

V. Transformation Techniques

A wide range of techniques for inserting exogenous polynucleotides are known for a number of host cells, including, without limitation, bacterial, yeast, mammalian, insect and plant

25 cells.

Techniques for transforming a wide variety of higher plant species are well known and described in the technical and scientific literature. See, e.g. Weising et al., *Ann. Rev. Genet.* 22:421 (1988); and Christou, *Euphytica*, v. 85, n.1-3:13-27, (1995).

DNA constructs of the invention may be introduced into the genome of the desired plant host by a variety of conventional techniques. For example, the DNA construct may be introduced directly into the genomic DNA of the plant cell using techniques such as electroporation and microinjection of plant cell protoplasts, or the DNA constructs can be introduced directly to plant tissue using ballistic methods, such as DNA particle bombardment. Alternatively, the DNA constructs may be combined with suitable T-DNA flanking regions and introduced into a conventional *Agrobacterium tumefaciens* host vector. The virulence functions of the *Agrobacterium tumefaciens* host will direct the insertion of the construct and adjacent marker into the plant cell DNA when the cell is infected by the bacteria (McCormac et al., *Mol. Biotechnol.* 8:199 (1997); Hamilton, *Gene* 200:107 (1997)); Salomon et al. *EMBO J.* 3:141 (1984); Herrera-Estrella et al. *EMBO J.* 2:987 (1983).

Microinjection techniques are known in the art and well described in the scientific and patent literature. The introduction of DNA constructs using polyethylene glycol precipitation is described in Paszkowski et al. *EMBO J.* 3:2717 (1984). Electroporation techniques are described in Fromm et al. *Proc. Natl Acad. Sci. USA* 82:5824 (1985). Ballistic transformation techniques are described in Klein et al. *Nature* 327:773 (1987). *Agrobacterium tumefaciens*-mediated transformation techniques, including disarming and use of binary or co-integrate vectors, are well described in the scientific literature. See, for example Hamilton, *CM., Gene* 200:107 (1997); Müller et al. *Mol. Gen. Genet.* 207:171 (1987); Komari et al. *Plant J.* 10:165 (1996); Venkateswarlu et al. *Biotechnology* 9:1103 (1991) and Gleave, *AP., Plant Mol. Biol.* 20:1203 (1992); Graves and Goldman, *Plant Mol. Biol.* 7:34 (1986) and Gould et al., *Plant Physiology* 95:426 (1991).

Transformed plant cells which are derived by any of the above transformation techniques can be cultured to regenerate a whole plant that possesses the transformed genotype and thus the desired phenotype such as seedlessness. Such regeneration techniques rely on manipulation of certain phytohormones in a tissue culture growth medium, typically relying on a biocide and/or herbicide marker which has been introduced together with the desired nucleotide sequences. Plant regeneration from cultured protoplasts is described in Evans et al., *Protoplasts Isolation and Culture* in "Handbook of Plant Cell Culture," pp. 124-176, MacMillan Publishing Company, New York, 1983; and Binding, *Regeneration of Plants, Plant Protoplasts*, pp. 21-73,

CRC Press, Boca Raton, 1988. Regeneration can also be obtained from plant callus, explants, organs, or parts thereof. Such regeneration techniques are described generally in Klee et al. *Ann. Rev. of Plant Phys.* 38:467 (1987). Regeneration of monocots (rice) is described by Hosoyama et al. (*Biosci. Biotechnol. Biochem.* 58:1500 (1994)) and by Ghosh et al. (*J. Biotechnol.* 32:1 (1994)). The nucleic acids of the invention can be used to confer desired traits on essentially any plant.

Thus, the invention has use over a broad range of plants, including species from the genera *Anacardium*, *Arachis*, *Asparagus*, *Atropa*, *Avena*, *Brassica*, *Citrus*, *Citrullus*, *Capsicum*, *Carthamus*, *Cocos*, *Coffea*, *Cucumis*, *Cucurbita*, *Daucus*, *Elaeis*, *Fragaria*, *Glycine*, *Gossypium*, *Helianthus*, *Heterocallis*, *Hordeum*, *Hyoscyamus*, *Lactuca*, *Linum*, *Lolium*, *Lupinus*, *Lycopersicon*, *Malus*, *Manihot*, *Majorana*, *Medicago*, *Nicotiana*, *Olea*, *Oryza*, *Panicum*, *Pannisetum*, *Persea*, *Phaseolus*, *Pistachia*, *Pisum*, *Pyrus*, *Prunus*, *Raphanus*, *Ricinus*, *Secale*, *Senecio*, *Sinapis*, *Solanum*, *Sorghum*, *Theobromus*, *Trigonella*, *Triticum*, *Vicia*, *Vitis*, *Vigna*, and, *Zea*.

One of skill will recognize that after the expression cassette is stably incorporated in transgenic plants and confirmed to be operable, it can be introduced into other plants by sexual crossing. Any of a number of standard breeding techniques can be used, depending upon the species to be crossed.

The particular sequences of SDFs identified are provided in the attached Tables 1 and 2. One of ordinary skill in the art, having this data, can obtain cloned DNA fragments, synthetic DNA fragments or polypeptides constituting desired sequences by recombinant methodology known in the art or described herein.

EXAMPLES

The invention is illustrated by way of the following examples. The invention is not limited by these examples as the scope of the invention is defined solely by the claims following.

EXAMPLE 1: cDNA PREPARATION

A number of the nucleotide sequences disclosed in Tables 1 and 2 herein as representative of the SDFs of the invention can be obtained by sequencing genomic DNA (gDNA) and/or cDNA from corn plants grown from HYBRID SEED # 35A19, purchased from

Pioneer Hi-Bred International, Inc., Supply Management, P.O. Box 256, Johnston, Iowa 50131-0256.

A number of the nucleotide sequences disclosed in Tables 1 and 2 herein as representative of the SDFs of the invention can also be obtained by sequencing genomic DNA from *Arabidopsis thaliana*, Wassilewskija ecotype or by sequencing cDNA obtained from mRNA from such plants as described below. This is a true breeding strain. Seeds of the plant are available from the Arabidopsis Biological Resource Center at the Ohio State University, under the accession number CS2360. Seeds of this plant were deposited under the terms and conditions of the Budapest Treaty at the American Type Culture Collection, Manassas, VA on August 31, 1999, and were assigned ATCC No. PTA-595.

Other methods for cloning full-length cDNA are described, for example, by Seki et al., *Plant Journal* 15:707-720 (1998) "High-efficiency cloning of Arabidopsis full-length cDNA by biotinylated Cap trapper"; Maruyama et al., *Gene* 138:171 (1994) "Oligo-capping a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides"; and WO 96/34981.

Tissues were, or each organ was, individually pulverized and frozen in liquid nitrogen. Next, the samples were homogenized in the presence of detergents and then centrifuged. The debris and nuclei were removed from the sample and more detergents were added to the sample. The sample was centrifuged and the debris was removed. Then the sample was applied to a 2M sucrose cushion to isolate polysomes. The RNA was isolated by treatment with detergents and proteinase K followed by ethanol precipitation and centrifugation. The polysomal RNA from the different tissues was pooled according to the following mass ratios: 15/15/1 for male inflorescences, female inflorescences and root, respectively. The pooled material was then used for cDNA synthesis by the methods described below.

Starting material for cDNA synthesis for the exemplary corn cDNA clones with sequences presented in Tables 1 and 2 was poly(A)-containing polysomal mRNAs from inflorescences and root tissues of corn plants grown from HYBRID SEED # 35A19. Male inflorescences and female (pre-and post-fertilization) inflorescences were isolated at various stages of development. Selection for poly(A) containing polysomal RNA was done using oligo d(T) cellulose columns, as described by Cox and Goldberg, "Plant Molecular Biology: A Practical Approach", pp. 1-35, Shaw ed., c. 1988 by IRL, Oxford. The quality and the integrity of the polyA+ RNAs were evaluated.

Starting material for cDNA synthesis for the exemplary *Arabidopsis* cDNA clones with sequences presented in Tables 1 and 2 was polysomal RNA isolated from the top-most inflorescence tissues of *Arabidopsis thaliana* Wassilewskija (Ws.) and from roots of *Arabidopsis thaliana* Landsberg erecta (L. er.), also obtained from the Arabidopsis Biological Resource Center. Nine parts inflorescence to every part root was used, as measured by wet mass. Tissue was pulverized and exposed to liquid nitrogen. Next, the sample was homogenized in the presence of detergents and then centrifuged. The debris and nuclei were removed from the sample and more detergents were added to the sample. The sample was centrifuged and the debris was removed and the sample was applied to a 2M sucrose cushion to isolate polysomal RNA. Cox et al., "Plant Molecular Biology: A Practical Approach", pp. 1-35, Shaw ed., c. 1988 by IRL, Oxford. The polysomal RNA was used for cDNA synthesis by the methods described below. Polysomal mRNA was then isolated as described above for corn cDNA. The quality of the RNA was assessed electrophoretically.

Following preparation of the mRNAs from various tissues as described above, selection of mRNA with intact 5' ends and specific attachment of an oligonucleotide tag to the 5' end of such mRNA was performed using either a chemical or enzymatic approach. Both techniques take advantage of the presence of the "cap" structure, which characterizes the 5' end of most intact mRNAs and which comprises a guanosine generally methylated once, at the 7 position.

The chemical modification approach involves the optional elimination of the 2', 3'-cis diol of the 3' terminal ribose, the oxidation of the 2', 3'-cis diol of the ribose linked to the cap of the 5' ends of the mRNAs into a dialdehyde, and the coupling of the such obtained dialdehyde to a derivatized oligonucleotide tag. Further detail regarding the chemical approaches for obtaining mRNAs having intact 5' ends are disclosed in International Application No.

WO96/34981 published November 7, 1996.

The enzymatic approach for ligating the oligonucleotide tag to the intact 5' ends of mRNAs involves the removal of the phosphate groups present on the 5' ends of uncapped incomplete mRNAs, the subsequent decapping of mRNAs having intact 5' ends and the ligation of the phosphate present at the 5' end of the decapped mRNA to an oligonucleotide tag. Further detail regarding the enzymatic approaches for obtaining mRNAs having intact 5' ends are disclosed in Dumas Milne Edwards J.B. (Doctoral Thesis of Paris VI University, Le clonage des ADNc complets: difficultés et perspectives nouvelles. Apports pour l'étude de la régulation de

l'expression de la tryptophane hydroxylase de rat, 20 Dec. 1993), EP0 625572 and Kato *et al.*, *Gene* 150:243-250 (1994).

In both the chemical and the enzymatic approach, the oligonucleotide tag has a restriction enzyme site (e.g. an EcoRI site) therein to facilitate later cloning procedures.

5 Following attachment of the oligonucleotide tag to the mRNA, the integrity of the mRNA is examined by performing a Northern blot using a probe complementary to the oligonucleotide tag.

For the mRNAs joined to oligonucleotide tags using either the chemical or the enzymatic method, first strand cDNA synthesis is performed using an oligo-dT primer with reverse
10 transcriptase. This oligo-dT primer can contain an internal tag of at least 4 nucleotides, which can be different from one mRNA preparation to another. Methylated dCTP is used for cDNA first strand synthesis to protect the internal EcoRI sites from digestion during subsequent steps. The first strand cDNA is precipitated using isopropanol after removal of RNA by alkaline hydrolysis to eliminate residual primers.

15 Second strand cDNA synthesis is conducted using a DNA polymerase, such as Klenow fragment and a primer corresponding to the 5' end of the ligated oligonucleotide. The primer is typically 20-25 bases in length. Methylated dCTP is used for second strand synthesis in order to protect internal EcoRI sites in the cDNA from digestion during the cloning process.

20 Following second strand synthesis, the full-length cDNAs are cloned into a phagemid vector, such as pBlueScript™ (Stratagene). The ends of the full-length cDNAs are blunted with T4 DNA polymerase (Biolabs) and the cDNA is digested with EcoRI. Since methylated dCTP is used during cDNA synthesis, the EcoRI site present in the tag is the only hemi-methylated site; hence the only site susceptible to EcoRI digestion. In some instances, to facilitate subcloning, an Hind III adapter is added to the 3' end of full-length cDNAs.

25 The full-length cDNAs are then size fractionated using either exclusion chromatography (AcA, Biosepra) or electrophoretic separation which yields 3 to 6 different fractions. The full-length cDNAs are then directionally cloned either into pBlueScript™ using either the EcoRI and SmaI restriction sites or, when the Hind III adapter is present in the full-length cDNAs, the EcoRI and Hind III restriction sites. The ligation mixture is transformed, preferably by
30 electroporation, into bacteria, which are then propagated under appropriate antibiotic selection.

Clones containing the oligonucleotide tag attached to full-length cDNAs are selected as follows.

The plasmid cDNA libraries made as described above are purified (e.g. by a column available from Qiagen). A positive selection of the tagged clones is performed as follows. Briefly, in this selection procedure, the plasmid DNA is converted to single stranded DNA using phage F1 gene II endonuclease in combination with an exonuclease (Chang et al., *Gene* 127:95 (1993)) such as exonuclease III or T7 gene 6 exonuclease. The resulting single stranded DNA is then purified using paramagnetic beads as described by Fry et al., *Biotechniques* 13: 124 (1992). Here the single stranded DNA is hybridized with a biotinylated oligonucleotide having a sequence corresponding to the 3' end of the oligonucleotide tag. Preferably, the primer has a length of 20-25 bases. Clones including a sequence complementary to the biotinylated oligonucleotide are selected by incubation with streptavidin coated magnetic beads followed by magnetic capture. After capture of the positive clones, the plasmid DNA is released from the magnetic beads and converted into double stranded DNA using a DNA polymerase such as ThermoSequenase™ (obtained from Amersham Pharmacia Biotech). Alternatively, protocols such as the Gene Trapper™ kit (Gibco BRL) can be used. The double stranded DNA is then transformed, preferably by electroporation, into bacteria. The percentage of positive clones having the 5' tag oligonucleotide is typically estimated to be between 90 and 98% from dot blot analysis.

Following transformation, the libraries are ordered in microtiter plates and sequenced. The *Arabidopsis* library was deposited at the American Type Culture Collection on January 7, 2000 as "*E-coli* liba 010600" under the accession number PTA-1161.

EXAMPLE 2: SOUTHERN HYBRIDIZATIONS

The SDFs of the invention can be used in Southern hybridizations as described above. The following describes extraction of DNA from nuclei of plant cells, digestion of the nuclear DNA and separation by length, transfer of the separated fragments to membranes, preparation of probes for hybridization, hybridization and detection of the hybridized probe.

The procedures described herein can be used to isolate related polynucleotides or for diagnostic purposes. Moderate stringency hybridization conditions, as defined above, are described in the present example. These conditions result in detection of hybridization between sequences having at least 70% sequence identity. As described above, the hybridization and wash conditions can be changed to reflect the desired percentatge of sequence identity between probe and target sequences that can be detected.

1031

In the following procedure, a probe for hybridization is produced from two PCR reactions using two primers from genomic sequence of *Arabidopsis thaliana*. As described above, the particular template for generating the probe can be any desired template.

The first PCR product is assessed to validate the size of the primer to assure it is of the expected size. Then the product of the first PCR is used as a template, with the same pair of primers used in the first PCR, in a second PCR that produces a labeled product used as the probe.

Fragments detected by hybridization, or other bands of interest, can be isolated from gels used to separate genomic DNA fragments by known methods for further purification and/or characterization.

Buffers for nuclear DNA extraction

1. 10X HB

	1000 ml	
40 mM spermidine	10.2 g	Spermine (Sigma S-2876) and spermidine (Sigma S-2501)
10 mM spermine	3.5 g	Stabilize chromatin and the nuclear membrane
0.1 M EDTA (disodium)	37.2 g	EDTA inhibits nuclease
0.1 M Tris	12.1 g	Buffer
0.8 M KCl	59.6 g	Adjusts ionic strength for stability of nuclei

Adjust pH to 9.5 with 10 N NaOH. It appears that there is a nuclease present in leaves. Use of pH 9.5 appears to inactivate this nuclease.

2. 2 M sucrose (684 g per 1000 ml)

1032

Heat about half the final volume of water to about 50°C. Add the sucrose slowly then bring the mixture to close to final volume; stir constantly until it has dissolved. Bring the solution to volume.

3. Sarkosyl solution (lyses nuclear membranes)

5		<u>1000 ml</u>
	N-lauroyl sarcosine (Sarkosyl)	20.0 g
	0.1 M Tris	12.1 g
	0.04 M EDTA (Disodium)	14.9 g

Adjust the pH to 9.5 after all the components are dissolved and bring up to the proper volume.

4. 20% Triton X-100
 80 ml Triton X-100
 320 ml 1xHB (w/o β -ME and PMSF)
 Prepare in advance; Triton takes some time to dissolve

A. Procedure

1. Prepare 1X "H" buffer (keep ice-cold during use)

		<u>1000 ml</u>
	10X HB	100 ml
	2 M sucrose	250 ml a non-ionic osmoticum
20	Water	634 ml

Added just before use:

100 mM PMSF*	10 ml a protease inhibitor; protects nuclear membrane proteins
β -mercaptoethanol	1 ml inactivates nuclease by reducing disulfide bonds

*100 mM PMSF

(phenyl methyl sulfonyl fluoride, Sigma P-7626)

(add 0.0875 g to 5 ml 100% ethanol)

2. Homogenize the tissue in a blender (use 300-400 ml of 1xHB per blender). Be sure that you use 5-10 ml of HB buffer per gram of tissue. Blenders generate heat so be sure to keep the homogenate cold. It is necessary to put the blenders in ice periodically.
3. Add the 20% Triton X-100 (25 ml per liter of homogenate) and gently stir on ice for 20 min. This lyses plastid, but not nuclear, membranes.
4. Filter the tissue suspension through several nylon filters into an ice-cold beaker. The first filtration is through a 250-micron membrane; the second is through an 85-micron membrane; the third is through a 50-micron membrane; and the fourth is through a 20-micron membrane. Use a large funnel to hold the filters. Filtration can be sped up by gently squeezing the liquid through the filters.
5. Centrifuge the filtrate at 1200 x g for 20 min. at 4°C to pellet the nuclei.
6. Discard the dark green supernatant. The pellet will have several layers to it. One is starch; it is white and gritty. The nuclei are gray and soft. In the early steps, there may be a dark green and somewhat viscous layer of chloroplasts.

Wash the pellets in about 25 ml cold H buffer (with Triton X-100) and resuspend by swirling gently and pipetting. After the pellets are resuspended.

Pellet the nuclei again at 1200 - 1300 x g. Discard the supernatant.

Repeat the wash 3-4 times until the supernatant has changed from a dark green to a pale green. This usually happens after 3 or 4 resuspensions. At this point, the pellet

1034

is typically grayish white and very slippery. The Triton X-100 in these repeated steps helps to destroy the chloroplasts and mitochondria that contaminate the prep.

Resuspend the nuclei for a final time in a total of 15 ml of H buffer and transfer the suspension to a sterile 125 ml Erlenmeyer flask.

- 5 7. Add 15 ml, dropwise, cold 2% Sarkosyl, 0.1 M Tris, 0.04 M EDTA solution (pH 9.5) while swirling gently. This lyses the nuclei. The solution will become very viscous.
8. Add 30 grams of CsCl and gently swirl at room temperature until the CsCl is in solution. The mixture will be gray, white and viscous.
9. Centrifuge the solution at 11,400 x g at 4°C for at least 30 min. The longer this spin is, the firmer the protein pellicle.
10. The result is typically a clear green supernatant over a white pellet, and (perhaps) under a protein pellicle. Carefully remove the solution under the protein pellicle and above the pellet. Determine the density of the solution by weighing 1 ml of solution and add CsCl if necessary to bring to 1.57 g/ml. The solution contains dissolved solids (sucrose etc) and the refractive index alone will not be an accurate guide to CsCl concentration.
11. Add 20 µl of 10 mg/ml EtBr per ml of solution.
12. Centrifuge at 184,000 x g for 16 to 20 hours in a fixed-angle rotor.
- 20 13. Remove the dark red supernatant that is at the top of the tube with a plastic transfer pipette and discard. Carefully remove the DNA band with another transfer pipette. The DNA band is usually visible in room light; otherwise, use a long wave UV light to locate the band.
- 25 14. Extract the ethidium bromide with isopropanol saturated with water and salt. Once the solution is clear, extract at least two more times to ensure that all of the EtBr is

gone. Be very gentle, as it is very easy to shear the DNA at this step. This extraction may take a while because the DNA solution tends to be very viscous. If the solution is too viscous, dilute it with TE.

15. Dialyze the DNA for at least two days against several changes (at least three times) of TE (10 mM Tris, 1mM EDTA, pH 8) to remove the cesium chloride.
16. Remove the dialyzed DNA from the tubing. If the dialyzed DNA solution contains a lot of debris, centrifuge the DNA solution at least at 2500 x g for 10 min. and carefully transfer the clear supernatant to a new tube. Read the A260 concentration of the DNA.
17. Assess the quality of the DNA by agarose gel electrophoresis (1% agarose gel) of the DNA. Load 50 ng and 100 ng (based on the OD reading) and compare it with known and good quality DNA. Undigested lambda DNA and a lambda-HindIII-digested DNA are good molecular weight makers.

Protocol for Digestion of Genomic DNA

Protocol:

1. The relative amounts of DNA for different crop plants that provide approximately a balanced number of genome equivalent is given in Table 3. Note that due to the size of the wheat genome, wheat DNA will be underrepresented. Lambda DNA provides a useful control for complete digestion.
2. Precipitate the DNA by adding 3 volumes of 100% ethanol. Incubate at -20°C for at least two hours. Yeast DNA can be purchased and made up at the necessary concentration, therefore no precipitation is necessary for yeast DNA.
3. Centrifuge the solution at 11,400 x g for 20 min. Decant the ethanol carefully (be careful not to disturb the pellet). Be sure that the residual ethanol is completely removed either by vacuum desiccation or by carefully wiping the sides of the tubes with a clean tissue.

4. Resuspend the pellet in an appropriate volume of water. Be sure the pellet is fully resuspended before proceeding to the next step. This may take about 30 min.
5. Add the appropriate volume of 10X reaction buffer provided by the manufacturer of the restriction enzyme to the resuspended DNA followed by the appropriate volume of enzymes. Be sure to mix it properly by slowly swirling the tubes.
6. Set-up the lambda digestion-control for each DNA that you are digesting.
7. Incubate both the experimental and lambda digests overnight at 37°C. Spin down condensation in a microfuge before proceeding.
8. After digestion, add 2 µl of loading dye (typically 0.25% bromophenol blue, 0.25% xylene cyanol in 15% Ficoll or 30% glycerol) to the lambda-control digests and load in 1% TPE-agarose gel (TPE is 90 mM Tris-phosphate, 2 mM EDTA, pH 8). If the lambda DNA in the lambda control digests are completely digested, proceed with the precipitation of the genomic DNA in the digests.
9. Precipitate the digested DNA by adding 3 volumes of 100% ethanol and incubating in -20°C for at least 2 hours (preferably overnight).

EXCEPTION: *Arabidopsis* and yeast DNA are digested in an appropriate volume; they don't have to be precipitated.

10. Resuspend the DNA in an appropriate volume of TE (e.g., 22 µl x 50 blots = 1100 µl) and an appropriate volume of 10X loading dye (e.g., 2.4 µl x 50 blots = 120 µl). Be careful in pipetting the loading dye - it is viscous. Be sure you are pipetting the correct volume.

Table 3

Some guide points in digesting genomic DNA.

			Genome	Amount
--	--	--	--------	--------

1037

Species	Genome Size	Size Relative to Arabidopsis	Equivalent to 2 μ g Arabidopsis DNA	of DNA per blot
Arabidopsis	120 Mb	1X	1X	2 μ g
Brassica	1,100 Mb	9.2X	0.54X	10 μ g
Corn	2,800 Mb	23.3X	0.43X	20 μ g
Cotton	2,300 Mb	19.2X	0.52X	20 μ g
Oat	11,300 Mb	94X	0.11X	20 μ g
Rice	400 Mb	3.3X	0.75X	5 μ g
Soybean	1,100 Mb	9.2X	0.54X	10 μ g
Sugarbeet	758 Mb	6.3X	0.8X	10 μ g
Sweetclover	1,100 Mb	9.2X	0.54X	10 μ g
Wheat	16,000 Mb	133X	0.08X	20 μ g
Yeast	15 Mb	0.12X	1X	0.25 μ g

Protocol for Southern Blot Analysis

The digested DNA samples are electrophoresed in 1% agarose gels in 1x TPE buffer. Low voltage; overnight separations are preferred. The gels are stained with EtBr and photographed.

1. For blotting the gels, first incubate the gel in 0.25 N HCl (with gentle shaking) for about 15 min.
2. Then briefly rinse with water. The DNA is denatured by 2 incubations. Incubate (with shaking) in 0.5 M NaOH in 1.5 M NaCl for 15 min.
3. The gel is then briefly rinsed in water and neutralized by incubating twice (with shaking) in 1.5 M Tris pH 7.5 in 1.5 M NaCl for 15 min.

1038

4. A nylon membrane is prepared by soaking it in water for at least 5 min, then in 6X SSC for at least 15 min. before use. (20x SSC is 175.3 g NaCl, 88.2 g sodium citrate per liter, adjusted to pH 7.0.)
5. The nylon membrane is placed on top of the gel and all bubbles in between are removed. The DNA is blotted from the gel to the membrane using an absorbent medium, such as paper toweling and 6x SCC buffer. After the transfer, the membrane may be lightly brushed with a gloved hand to remove any agarose sticking to the surface.
6. The DNA is then fixed to the membrane by UV crosslinking and baking at 80°C. The membrane is stored at 4°C until use.

B. Protocol for PCR Amplification of Genomic Fragments in Arabidopsis

Amplification procedures:

1. Mix the following in a 0.20 ml PCR tube or 96-well PCR plate:

Volume	Stock	Final Amount or Conc.
0.5 μ l	~ 10 ng/ μ l genomic DNA ¹	5 ng
2.5 μ l	10X PCR buffer	20 mM Tris, 50 mM KCl
0.75 μ l	50 mM MgCl ₂	1.5 mM
1 μ l	10 pmol/ μ l Primer 1 (Forward)	10 pmol
1 μ l	10 pmol/ μ l Primer 2 (Reverse)	10 pmol
0.5 μ l	5 mM dNTPs	0.1 mM

¹ Arabidopsis DNA is used in the present experiment, but the procedure is a general one.

1039

0.1 μ l	5 units/ μ l Platinum Taq™ (Life Technologies, Gaithersburg, MD) DNA Polymerase	1 units
(to 25 μ l)	Water	

2. The template DNA is amplified using a Perkin Elmer 9700 PCR machine:

1) 94°C for 10 min. followed by

<u>2)</u> 5 cycles:	<u>3)</u> 5 cycles:	<u>4)</u> 25 cycles:
94 °C - 30 sec 62 °C - 30 sec 72 °C - 3 min	94 °C - 30 sec 58 °C - 30 sec 72 °C - 3 min	94 °C - 30 sec 53 °C - 30 sec 72 °C - 3 min

5) 72°C for 7 min. Then the reactions are stopped by chilling to 4°C.

The procedure can be adapted to a multi-well format if necessary.

5 Quantification and Dilution of PCR Products:

1. The product of the PCR is analyzed by electrophoresis in a 1% agarose gel. A linearized plasmid DNA can be used as a quantification standard (usually at 50, 100, 200, and 400 ng). These will be used as references to approximate the amount of PCR products. HindIII-digested Lambda DNA is useful as a molecular weight marker. The gel can be run fairly quickly; e.g., at 100 volts. The standard gel is examined to determine that the size of the PCR products is consistent with the expected size and if there are significant extra bands or smeary products in the PCR reactions.

2. The amounts of PCR products can be estimated on the basis of the plasmid standard.

3. For the small number of reactions that produce extraneous bands, a small amount of DNA from bands with the correct size can be isolated by dipping a sterile 10- μ l tip into the band while viewing through a UV Transilluminator. The small amount of agarose gel (with the DNA fragment) is used in the labeling reaction.

5 C. Protocol for PCR-DIG-Labeling of DNA

Solutions:

Reagents in PCR reactions (diluted PCR products, 10X PCR Buffer, 50 mM MgCl₂, 5 U/ μ l Platinum Taq Polymerase, and the primers)

10X dNTP + DIG-11-dUTP [1:5]: (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.65 mM dTTP, 0.35 mM DIG-11-dUTP)

10X dNTP + DIG-11-dUTP [1:10]: (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.81 mM dTTP, 0.19 mM DIG-11-dUTP)

10X dNTP + DIG-11-dUTP [1:15]: (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.875 mM dTTP, 0.125 mM DIG-11-dUTP)

TE buffer (10 mM Tris, 1 mM EDTA, pH 8)

Maleate buffer: In 700 ml of deionized distilled water, dissolve 11.61 g maleic acid and 8.77 g NaCl. Add NaOH to adjust the pH to 7.5. Bring the volume to 1 L. Stir for 15 min. and sterilize.

10% blocking solution: In 80 ml deionized distilled water, dissolve 1.16g maleic acid. Next, add NaOH to adjust the pH to 7.5. Add 10 g of the blocking reagent powder (Boehringer Mannheim, Indianapolis, IN, Cat. no. 1096176). Heat to 60°C while stirring to dissolve the powder. Adjust the volume to 100 ml with water. Stir and sterilize.

1% blocking solution: Dilute the 10% stock to 1% using the maleate buffer.

Buffer 3 (100 mM Tris, 100 mM NaCl, 50 mM MgCl₂, pH9.5). Prepared from autoclaved solutions of 1M Tris pH 9.5, 5 M NaCl, and 1 M MgCl₂ in autoclaved distilled water.

1000 900 800 700 600 500 400 300 200 100 0

Procedure:

1. PCR reactions are performed in 25 µl volumes containing:

PCR buffer	1X
MgCl ₂	1.5 mM
5 10X dNTP + DIG-11-dUTP	1X (please see the note below)
Platinum Taq™ Polymerase	1 unit
10 pg probe DNA	
10 pmol primer 1	

Note:Use for:

10X dNTP + DIG-11-dUTP (1:5)	< 1 kb
10X dNTP + DIG-11-dUTP (1:10)	1 kb to 1.8 kb
10X dNTP + DIG-11-dUTP (1:15)	> 1.8 kb

2. The PCR reaction uses the following amplification cycles:

- 1) 94°C for 10 min.

<u>2)</u> 5 cycles:	<u>3)</u> 5 cycles:	<u>4)</u> 25 cycles:
95°C - 30 sec	95°C - 30 sec	95°C - 30 sec
61°C - 1 min	59°C - 1 min	51°C - 1 min
73°C - 5 min	75°C - 5 min	73°C - 5 min

- 5) 72°C for 8 min. The reactions are terminated by chilling to 4°C (hold).

3. The products are analyzed by electrophoresis- in a 1% agarose gel, comparing to an aliquot of the unlabelled probe starting material.
4. The amount of DIG-labeled probe is determined as follows:

Make serial dilutions of the diluted control DNA in dilution buffer (TE: 10 mM Tris and 1 mM EDTA, pH 8) as shown in the following table:

DIG-labeled control DNA starting conc.	Stepwise Dilution	Final Conc. (Dilution Name)
5 ng/μl	1 μl in 49 μl TE	100 pg/μl (A)
100 pg/μl (A)	25 μl in 25 μl TE	50 pg/μl (B)
50 pg/μl (B)	25 μl in 25 μl TE	25 pg/μl (C)
25 pg/μl (C)	20 μl in 30 μl TE	10 pg/μl (D)

- a. Serial dilutions of a DIG-labeled standard DNA ranging from 100 pg to 10 pg are spotted onto a positively charged nylon membrane, marking the membrane lightly with a pencil to identify each dilution.
- b. Serial dilutions (e.g., 1:50, 1:2500, 1:10,000) of the newly labeled DNA probe are spotted.
- c. The membrane is fixed by UV crosslinking.
- d. The membrane is wetted with a small amount of maleate buffer and then incubated in 1% blocking solution for 15 min at room temp.
- e. The labeled DNA is then detected using alkaline phosphatase conjugated anti-DIG antibody (Boehringer Mannheim, Indianapolis, IN, cat. no. 1093274) and an NBT substrate according to the manufacture's instruction.
- f. Spot intensities of the control and experimental dilutions are then compared to estimate the concentration of the PCR-DIG-labeled probe.

D. Prehybridization and Hybridization of Southern BlotsSolutions:

100% Formamide purchased from Gibco

20X SSC (1X = 0.15 M NaCl, 0.015 M Na₃citrate)

per L: 175 g NaCl
87.5 g Na₃citrate·2H₂O

20% Sarkosyl (N-lauroyl-sarcosine)

20% SDS (sodium dodecyl sulphate)

10% Blocking Reagent: In 80 ml deionized distilled water, dissolve 1.16 g maleic acid. Next, add NaOH to adjust the pH to 7.5. Add 10 g of the blocking reagent powder. Heat to 60°C while stirring to dissolve the powder. Adjust the volume to 100 ml with water. Stir and sterilize.

Prehybridization Mix:

Final Concentration	Components	Volume (per 100 ml)	Stock
50%	Formamide	50 ml	100%
5X	SSC	25 ml	20X
0.1%	Sarkosyl	0.5 ml	20%
0.02%	SDS	0.1 ml	20%
2%	Blocking Reagent	20 ml	10%
	Water	4.4 ml	

General Procedures:

1. Place the blot in a heat-sealable plastic bag and add an appropriate volume of prehybridization solution (30 ml/100cm²) at room temperature. Seal the bag with a heat sealer, avoiding bubbles as much as possible. Lay down the bags in a large plastic tray (one tray can accommodate at least 4–5 bags). Ensure that the bags are

1045

lying flat in the tray so that the prehybridization solution is evenly distributed throughout the bag. Incubate the blot for at least 2 hours with gentle agitation using a waver shaker.

2. Denature DIG-labeled DNA probe by incubating for 10 min. at 98°C using the PCR machine and immediately cool it to 4°C.

3. Add probe to prehybridization solution (25 ng/ml; 30 ml = 750 ng total probe) and mix well but avoid foaming. Bubbles may lead to background.

4. Pour off the prehybridization solution from the hybridization bags and add new prehybridization and probe solution mixture to the bags containing the membrane.

5. Incubate with gentle agitation for at least 16 hours.

6. Proceed to medium stringency post-hybridization wash:

Three times for 20 min. each with gentle agitation using 1X SSC, 1% SDS at 60°C.

All wash solutions must be prewarmed to 60°C. Use about 100 ml of wash solution per membrane.

To avoid background keep the membranes fully submerged to avoid drying in spots; agitate sufficiently to avoid having membranes stick to one another.

7. After the wash, proceed to immunological detection and CSPD development.

E. Procedure for Immunological Detection with CSPD

Solutions:

Buffer 1: Maleic acid buffer (0.1 M maleic acid, 0.15 M NaCl; adjusted to pH 7.5 with NaOH)

1046

Washing buffer: Maleic acid buffer with 0.3% (v/v) Tween 20.

Blocking stock solution 10% blocking reagent in buffer 1. Dissolve (10X concentration): blocking reagent powder (Boehringer Mannheim, Indianapolis, IN, cat. no. 1096176) by constantly stirring on a 65°C heating block or heat in a microwave, autoclave and store at 4°C.

Buffer 2

(1X blocking solution): Dilute the stock solution 1:10 in Buffer 1.

Detection buffer: 0.1 M Tris, 0.1 M NaCl, pH 9.5

Procedure:

1. After the post-hybridization wash the blots are briefly rinsed (1-5 min.) in the maleate washing buffer with gentle shaking.
2. Then the membranes are incubated for 30 min. in Buffer 2 with gentle shaking.
3. Anti-DIG-AP conjugate (Boehringer Mannheim, Indianapolis, IN, cat. no. 1093274) at 75 mU/ml (1:10,000) in Buffer 2 is used for detection. 75 ml of solution can be used for 3 blots.
4. The membrane is incubated for 30 min. in the antibody solution with gentle shaking.
5. The membrane are washed twice in washing buffer with gentle shaking. About 250 mls is used per wash for 3 blots.
6. The blots are equilibrated for 2-5 min in 60 ml detection buffer.
7. Dilute CSPD (1:200) in detection buffer. (This can be prepared ahead of time and stored in the dark at 4°C).

The following steps must be done individually. Bags (one for detection and one for exposure) are generally cut and ready before doing the following steps.

8. The blot is carefully removed from the detection buffer and excess liquid removed without drying the membrane. The blot is immediately placed in a bag and 1.5 ml of CSPD solution is added. The CSPD solution can be spread over the membrane. Bubbles present at the edge and on the surface of the blot are typically removed by gentle rubbing. The membrane is incubated for 5 min. in CSPD solution.
9. Excess liquid is removed and the membrane is blotted briefly (DNA side up) on Whatman 3MM paper. Do not let the membrane dry completely.
10. Seal the damp membrane in a hybridization bag and incubate for 10 min at 37°C to enhance the luminescent reaction.
11. Expose for 2 hours at room temperature to X-ray film. Multiple exposures can be taken. Luminescence continues for at least 24 hours and signal intensity increases during the first hours.

Example 3: Transformation of Carrot Cells

Transformation of plant cells can be accomplished by a number of methods, as described above. Similarly, a number of plant genera can be regenerated from tissue culture following transformation. Transformation and regeneration of carrot cells as described herein is illustrative.

Single cell suspension cultures of carrot (*Daucus carota*) cells are established from hypocotyls of cultivar Early Nantes in B₅ growth medium (O.L. Gamborg et al., *Plant Physiol.* 45:372 (1970)) plus 2,4-D and 15 mM CaCl₂ (B₅-44 medium) by methods known in the art. The suspension cultures are subcultured by adding 10 ml of the suspension culture to 40 ml of B₅-44 medium in 250 ml flasks every 7 days and are maintained in a shaker at 150 rpm at 27 °C in the dark.

The suspension culture cells are transformed with exogenous DNA as described by Z. Chen et al. *Plant Mol. Bio.* 36:163 (1998). Briefly, 4-days post-subculture cells are incubated with cell wall digestion solution containing 0.4 M sorbitol, 2% driselase, 5mM MES (2-[N-

Morpholino] ethanesulfonic acid) pH 5.0 for 5 hours. The digested cells are pelleted gently at 60 xg for 5 min. and washed twice in W5 solution containing 154 mM NaCl, 5 mM KCl, 125 mM CaCl₂ and 5mM glucose, pH 6.0. The protoplasts are suspended in MC solution containing 5 mM MES, 20 mM CaCl₂, 0.5 M mannitol, pH 5.7 and the protoplast density is adjusted to about 4×10^6 protoplasts per ml.

15-60 µg of plasmid DNA is mixed with 0.9 ml of protoplasts. The resulting suspension is mixed with 40% polyethylene glycol (MW 8000, PEG 8000), by gentle inversion a few times at room temperature for 5 to 25 min. Protoplast culture medium known in the art is added into the PEG-DNA-protoplast mixture. Protoplasts are incubated in the culture medium for 24 hour to 5 days and cell extracts can be used for assay of transient expression of the introduced gene. Alternatively, transformed cells can be used to produce transgenic callus, which in turn can be used to produce transgenic plants, by methods known in the art. See, for example, Nomura and Komamine, *Plt. Phys.* 79:988-991 (1985), *Identification and Isolation of Single Cells that Produce Somatic Embryos in Carrot Suspension Cultures*.

An additional deposit of an *E. coli* Library, *E. coli*LibA021800, was made at the American Type Culture Collection in Manassas, Virginia, USA on February 22, 2000 to meet the requirements of Budapest Treaty for the international recognition of the deposit of microorganisms.

The invention being thus described, it will be apparent to one of ordinary skill in the art that various modifications of the materials and methods for practicing the invention can be made. Such modifications are to be considered within the scope of the invention as defined by the following claims.

Each of the references from the patent and periodical literature cited herein is hereby expressly incorporated in its entirety by such citation.